Pós-Graduação em Ciência da Computação

Hilário Tomaz Alves de Oliveira

Sumarização Automática de Textos Baseada em Conceitos via Programação Linear Inteira e Regressão

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
http://cin.ufpe.br/~posgraduacao

RECIFE

2018

Hilário Tomaz Alves de Oliveira

Sumarização Automática de Textos Baseada em Conceitos via Programação Linear Inteira e Regressão

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro de Informática, da Universidade Federal de Pernambuco, como requisito para a obtenção do título de Doutor em Ciência da Computação.

Orientador: Dr. Frederico Luiz Gonçalves de Freitas

Coorientador: Dr. Rinaldo José de Lima

RECIFE
2018

Catalogação na fonte
Bibliotecário Jefferson Luiz Alves Nazareno CRB4-1758

O48s Oliveira, Hilário Tomaz Alves de.
Sumarização automática de textos baseada em conceitos via programação linear inteira e regressão / Hilário Tomaz Alves de Oliveira. – 2018.
216 f.: fig. tab.

Orientador: Frederico Luiz Gonçalves de Freitas.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn. Ciência da Computação, Recife, 2018.
Inclui referências e apêndices.

1. Inteligência artificial. 2. Processamento de linguagem natural . 3. Mineração de texto. 4. Sumarização automática de texto. I. Freitas, Frederico Luiz Gonçalves. (Orientador) II. Titulo.

006.3 CDD (22. ed.) UFPE-MEI 2018- 66

Hilário Tomaz Alves de Oliveira

Sumarização Automática de Textos Baseada em Conceitos via Programação Linear Inteira e Regressão

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro de Informática, da Universidade Federal de Pernambuco, como requisito para a obtenção do título de Doutor em Ciência da Computação.

Aprovado em: 08/03/2018

Prof. Orientador: Dr. Frederico Luiz Gonçalves de Freitas

BANCA EXAMINADORA

Profa. Dra. Flávia de Almeida Barros
Centro de Informática/UFPE

Profa. Dra. Renata Maria Cardoso Rodrigues de Souza
Centro de Informática/UFPE

Prof. Dr. Thiago Alexandre Salgueiro Pardo
Instituto de Ciências Matemáticas e de Computação/USP

Profa. Dra. Daniela Barreiro Claro
Departamento de Ciência da Computação/UFBA

Prof. Dr. Renato Fernandes Corrêa
Departamento de Ciência da Informação/UFPE

Decido este trabalho à minha família que sempre foi fonte de inspiração e meu porto seguro perante os desafios enfrentados durante toda minha vida.

AGRADECIMENTOS

Agradeço à minha família, especialmente meus pais (Francisco Tomaz e Luzineida Ferreira) e meus irmãos (Daniel Tomaz e Danielle Tomaz), por sempre me apoiarem incondicionalmente e por me incentivarem a nunca desistir dos meus sonhos.

Ao meu orientador, professor Dr. Fred Freitas, pela confiança, competência, palavras de incentivo, dedicação e orientação desde o mestrado.

Ao meu coorientador, professor Dr. Rinaldo Lima pelo suporte, incentivo e pelas discussões que ajudaram muito o desenvolvimento deste trabalho.

Aos membros do grupo de pesquisa em Sumarização Automática de Textos, em especial ao professor Dr. Rafael Dueire Lins, pelas contribuições e grandes ensinamentos.

Aos meus amigos de convívio no apartamento Renê Gadelha, Thomas Cristanis e Edigleison Barbosa, pela convivência ao longo desses quatro anos de doutorado, cujos momentos lembrarei por toda a minha vida.

A todos os meus amigos, em especial João Emanoel, Alex Nery, Jamilson, Francisco Airton e Adriano Ferraz por todos os momentos de descontração que ajudaram e muito a esquecer um pouco dos desafios do dia a dia de um curso de doutorado e também de morar longe da família.

Aos amigos do laboratório da Associação de Pós-graduação (APG), pela convivência durante os quatro anos de doutorado e pelas conversas, sempre interessantes, durante os intervalos do café.

Aos membros da banca examinadora pelas contribuições e direcionamentos no intuito de enriquecer este trabalho.

Ao Centro de Informática da Universidade Federal de Pernambuco - CIN-UFPE, pela excelente estrutura física e pessoal proporcionada a todos os seus alunos.

Por fim, mas não menos importante, ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo auxílio financeiro.

"Pouco conhecimento faz com que as pessoas se sintam orgulhosas. Muito conhecimento, que se sintam humildes. É assim que as espigas sem grãos erguem desdenhosamente a cabeça para o Céu, enquanto que as cheias as baixam para a terra, sua mãe."

(Leonardo da Vinci)

RESUMO

Sumarização Automática de Textos é o processo de criação automático de um resumo contendo as informações mais relevantes, a partir de um único documento (monodocumento) ou de um grupo de documentos relacionados (multidocumento). O interesse no desenvolvimento de novos e eficientes sistemas de sumarização é crescente, já que eles possuem o potencial de auxiliar no processamento de grandes volumes de documentos textuais, ressaltando as informações mais relevantes para os usuários. Apesar dos avanços obtidos nos últimos anos, ainda existe uma grande diferença entre os resumos gerados automaticamente e os escritos por seres humanos. A maioria das atuais estratégias de sumarização são estáticas, ou seja, adotam um método de sumarização com um conjunto de parâmetros pré-definido para todos os documentos de entrada. Investigações recentes na literatura e experimentos conduzidos neste trabalho demonstram que essa característica é uma significante limitação, já que a adoção de um único método de sumarização não consegue obter um alto desempenho para todos os documentos, mesmo quando eles pertencem ao mesmo domínio. Neste contexto, este trabalho propõe uma abordagem baseada em conceitos utilizando Programação Linear Inteira (PLI) e regressão para as tarefas de sumarização monodocumento e multidocumento de artigos de notícias. A arquitetura da solução proposta é composta por duas etapas principais: a geração de diversos resumos candidatos e, posteriormente, a identificação e seleção do resumo mais informativo. Duas abordagens baseadas em conceitos usando PLI são propostas para a geração dos resumos candidatos nas tarefas de sumarização monodocumento e multidocumento. Tais abordagens possibilitam a exploração de diversas configurações, permitindo a geração de uma grande variedade de resumos candidatos representando diferentes perspectivas. As abordagens propostas são integradas em uma macro arquitetura com a etapa de seleção do resumo mais informativo. Essa etapa de seleção vislumbra estimar a cobertura de informações relevantes presentes nos resumos candidatos gerados, permitindo assim, a identificação do resumo estimado como mais representativo. Para isso, diversas características baseadas em tradicionais indicadores de relevância, como posição, frequência e centralidade, identificadas na literatura e outras propostas neste trabalho, são exploradas para a construção de um modelo de regressão. Diversos experimentos foram conduzidos nos principais corpora da área, visando avaliar diferentes aspectos das abordagens propostas nas tarefas de sumarização monodocumento e multidocumento. Os resultados obtidos demonstram que as soluções propostas, para ambas as tarefas de sumarização, são capazes de aumentar a informatividade dos resumos gerados, com base nas medidas de cobertura do ROUGE-1 e ROUGE-2, em comparação com outros sistemas do estado da arte.

Palavras-chave: Sumarização Automática de Textos. Sumarização Extrativa Monodocumento. Sumarização Extrativa Multidocumento. Programação Linear Inteira. Regressão.

ABSTRACT

Automatic Text Summarization (ATS) is the process of automatically creating a summary containing the most relevant information from a unique document (single-document) or a group of related documents (multi-document). The interest in developing new and efficient summarization systems is increasing, since they have the potential to assist the processing of large volumes of textual documents, highlighting the most relevant information for users. Despite the advances achieved in recent years, there is still a considerable difference between automatically generated summaries and those written by human beings. Most current summarization approaches are static, i.e., they adopt a summarization method with a predefined set of parameters for all input documents. Recent investigations in the literature and experiments conducted in this work demonstrate that this characteristic is a significant limitation since the adoption of a single summarization method cannot obtain high performance for all documents, even when they belong to the same domain. In this context, this work proposes a concept-based approach, employing Integer Linear Programming (ILP) and regression for single- and multi-document summarization of news articles. The architecture of the proposed solution consists of two main steps: the generation of several candidate summaries and, later, the identification and selection of the most informative summary. Two concept-based ILP approaches are proposed for the generation of candidate summaries in the single- and multi-document summarization tasks. Such approaches enable the exploration of several configurations, allowing the generation of a large variety of candidate summaries representing different perspectives. The proposed approaches are integrated into a macro-architecture with the most informative summary selection step. This selection stage envisages estimating the coverage of relevant information present in the candidate summaries generated, allowing the identification of the candidate summary estimated as the most informative. Several characteristics based on traditional content importance indicators, such as position, frequency, and centrality, identified in the literature and other proposed in this work, are explored for the construction of a regression model. Several experiments were conducted in the most adopted corpora of the area aiming to evaluate different aspects of the proposed approaches in the tasks of single- and multi-document summarization. The experimental results show that the proposed approaches, for both summarization tasks, can increase the informativeness of the generated summaries, based on the recall measures of ROUGE-1 and ROUGE-2, compared to other state-of-the-art systems.

Keywords: Automatic Text Summarization. Extractive Single-document Summarization. Extractive Multi-document Summarization. Integer Linear Programming.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

Alvin Toffler, em 1970, criou a expressão "Sobrecarga de Informação", do inglês *Information Overload*, quando ele previu que o aumento exponencial na produção de informações, eventualmente, traria dificuldades para as pessoas tomarem decisões com base na imensa quantidade de dados disponíveis (TOFFLER, 1970). Tal cenário é a realidade atualmente. A *Web*, por exemplo, possibilita criar, compartilhar e acessar uma vasta quantidade de informações digitais, especialmente documentos textuais, como artigos de notícias, livros, blogs, e-mails, artigos científicos, postagens em redes sociais, entre outros. Apesar das constantes evoluções no desenvolvimento dos motores de busca na *Web*, identificar informações úteis a partir dessa enorme quantidade de dados ainda é uma tarefa difícil e, em certos casos, inviável de ser realizada manualmente. Nesse contexto, existe um constante interesse no desenvolvimento de ferramentas capazes de recuperar, classificar e sintetizar informações relevantes de maneira rápida e eficiente.

Neste cenário, a Sumarização Automática de Textos (SAT) surge como uma possível solução para reduzir o tempo dos usuários em identificar as informações mais relevantes de um conjunto de documentos textuais. A SAT pode ser definida como a tarefa de geração automática de uma versão condensada (resumo), a partir de um único documento (monodocumento) ou de uma coleção de documentos relacionados (multidocumento), mantendo somente as informações mais relevantes (NENKOVA; MCKEOWN, 2012). De acordo com o dicionário Cambridge[1], um resumo pode ser definido como "uma concisa e breve descrição que apresenta os principais fatos ou ideias sobre alguma coisa". Apesar de o foco desta pesquisa de doutorado ser na sumarização de documentos textuais, existem diversos outros trabalhos que realizam esse processo analisando outros tipos de documentos, como vídeos (GUO et al., 2016) ou voz (KOZIEL et al., 2015).

Resumos têm desempenhado um papel importante para a sociedade, principalmente auxiliando os leitores a decidirem se determinado conteúdo satisfaz ou não os seus interesses. Os primeiros resumos de que se tem conhecimento foram criados em Roma por Plínio "o Velho", no ano de setenta e sete depois de Cristo (WIEGAND; JR., 1994). Plínio foi o autor da enciclopédia História Natural (*Naturalis Historia*), um vasto relato das ciências antigas distribuído em trinta e seis capítulos. Para facilitar a leitura e poupar tempo do imperador Tito Flávio Vespasiano Augusto, Plínio escreveu um resumo para cada um dos capítulos explicando quais eram os principais assuntos abordados.

Atualmente, com a evolução da tecnologia, e principalmente da *Internet*, a importância dos resumos se tornou ainda maior, sendo possível encontrá-los em diversos cenários. Alguns exemplos de aplicações nas quais resumos podem ser encontrados no cotidiano

[1] http://dictionary.cambridge.org/us/

são: as manchetes de jornais; os resumos de livros e artigos científicos; as revisões de opiniões sobre filmes, serviços ou produtos; as prévias (*snippets*) do conteúdo das páginas resultantes de uma pesquisa em um mecanismo de busca na *Web*; entre outras.

Apesar dos resumos serem adotados com frequência em diversos cenários, a maioria desses resumos são criados manualmente. Diante do crescente aumento na produção de informações, existem oportunidades para diversas novas aplicações, principalmente com o desenvolvimento de sistemas capazes de gerar resumos de qualidade de forma automática. Por exemplo, dada a vasta disponibilidade de notícias de fontes diversas, a existência de resumos que descrevessem as principais informações de cada notícia ou de um grupo de notícias correlacionadas seria de grande auxílio para os leitores. Atualmente, ainda é escasso o número de portais de notícias que disponibilizam resumos descrevendo os principais assuntos dos seus artigos de notícias. Além disso, serviços que agregam notícias de várias fontes, como o *Google News*[2], não fornecem um resumo único que sintetize os principais fatos de um grupo de notícias relacionadas sobre o mesmo evento ou assunto. Diante do exposto, o foco deste trabalho é na sumarização de artigos de notícias.

Cada uma das aplicações mencionadas anteriormente possui características diferentes, tais como o domínio e o tamanho dos documentos de entrada, o tipo de resumo a ser gerado, entre outras. Contudo, uma característica fundamental que os resumos devem ter para que sejam úteis para todas essas aplicações é que eles devem ser uma versão reduzida do documento original, mantendo-se relevante para atender aos requisitos dos usuários, e principalmente mantendo as informações mais relevantes dos documentos originais. Além disso, como os resumos gerados são destinados para leitura humana, eles devem ser coesos, coerentes e gramaticalmente corretos.

1.1 Definição do Problema

O processo de sumarização manual, realizada por seres humanos, é uma tarefa complexa e desafiadora para ser automatizada, pois requer habilidades e conhecimentos a priori que são difíceis de serem modelados computacionalmente (TORRES-MORENO, 2014). Os seres humanos exercitam esse processo todos os dias, mesmo que de forma desapercebida (por exemplo, quando se resume as informações de uma notícia, ou simplesmente contando a alguém como foi o seu dia no trabalho). Esse processo é muito suscetível à subjetividade, por exemplo, duas ou mais pessoas podem gerar resumos distintos e adequados para um mesmo evento ou documento. Isso acontece porque não é uma tarefa trivial identificar o que é ou não relevante para ser incluído no resumo a ser gerado. Além disso, a noção de relevância pode variar de uma pessoa para outra.

As primeiras pesquisas envolvendo a automatização do processo de sumarização datam de 1958 com o trabalho de Luhn (LUHN, 1958). Apesar disso, um grande interesse e

[2] https://news.google.com.br/

progresso na área somente aconteceram a partir da década de 1990 (TORRES-MORENO, 2014). Diversos métodos de sumarização têm sido propostos e avaliados desde então. Esses métodos variam desde a aplicação de técnicas para pontuar a relevâncias das frases de um documento (por exemplo, com base em sua posição, similaridade com o título do documento); a técnicas estatísticas, como frequência e coocorrência das palavras. Outros métodos mais sofisticados também têm sido investigados, entre eles se destacam métodos baseados em grafos (MIHALCEA; TARAU, 2004), algoritmos de agrupamento (WAN; YANG, 2008), métodos baseados em otimização combinatória utilizando Programação Linear Inteira (PLI) (GILLICK et al., 2009; LI; QIAN; LIU, 2013; BOUDIN; MOUGARD; FAVRE, 2015; LI; LIU; ZHAO, 2015), algoritmos de aprendizagem de máquina supervisionados (FATTAH, 2014; SILVA et al., 2015a; HONG; MARCUS; NENKOVA, 2015; WAN et al., 2015), entre outros.

Em geral, os métodos de SAT podem ser classificados em duas grandes abordagens: *Extrativa* e *Abstrativa* (NENKOVA; MCKEOWN, 2012). Abordagens extrativas identificam e selecionam o subconjunto de frases mais relevantes de um ou mais documentos, e as utilizam sem nenhuma alteração para a criação do resumo. Esse tipo de resumo, por utilizar as sentenças originais sem qualquer tipo de alteração, geralmente apresenta problemas de coesão (CHRISTENSEN et al., 2013) (por exemplo, correferências em aberto e quebras no fluxo de ideias entre as frases). Abordagens abstrativas buscam suprir essas limitações adotando técnicas que tentam simular a forma com que os seres humanos produzem resumos. Essa abordagem foca na seleção das informações mais relevantes dos documentos, e como expressá-las de uma nova forma, usando métodos para compressão de sentenças (ZAJIC et al., 2007; GILLICK et al., 2009), fusão de sentenças (FILIPPOVA, 2010; BANERJEE; MITRA; SUGIYAMA, 2015b), geração de novas frases usando linguagem natural (KHAN; SALIM; KUMAR, 2015), entre outras. Apesar de potencialmente gerarem resumos com maior qualidade do que as abordagens extrativas, muitas das técnicas adotadas em abordagens abstrativas ainda não estão desenvolvidas o suficiente para serem aplicadas em um domínio muito amplo, como é caso de artigos de notícias (SAGGION; POIBEAU, 2013). Diante desses fatos, as abordagens extrativas ainda são promissoras e mais exploradas atualmente, sendo, portanto, o foco deste trabalho.

Geralmente, as abordagens extrativas são executadas em três etapas (NENKOVA; MCKEOWN, 2012): **(i)** Criação de uma representação intermediária; **(ii)** Mensuração da importância das sentenças; e **(iii)** Geração de resumo. Documentos textuais estão em um formato não estruturado; assim, é necessário pré-processar e representar esses documentos de uma forma estruturada. A primeira etapa envolve geralmente algumas tarefas de Processamento de Linguagem Natural (PLN), como dividir o texto em parágrafos, frases, palavras, remoção de termos muito frequentes (*stopwords*), entre outras. Métodos para representar os principais temas discutidos no documento também são utilizados. Tais métodos podem computar a frequência ou a coocorrência das palavras, identificar a localização das sentenças no documento, verificar a presença de expressões chaves, entre

outras. Essa etapa também visa estimar quais são as frases mais relevantes, com base na representação criada anteriormente. Para isso, cada frase do documento recebe uma pontuação, como uma medida de sua relevância. Finalmente, na terceira etapa, as sentenças com maior pontuação são selecionadas para compor o resumo final. Uma das questões fundamentais nessa etapa é evitar a inclusão de redundância no resumo gerado, isto é, sentenças com um alto grau de sobreposição de informações.

Portanto, o grande desafio de qualquer sistema de SAT extrativo é lidar com duas questões fundamentais (SAGGION; POIBEAU, 2013): **(i)** Como identificar e selecionar as sentenças mais relevantes de um ou mais documentos; e **(ii)** Como organizá-las no resumo gerado, garantindo a qualidade linguística e evitando que informações redundantes sejam inseridas.

1.2 Lacunas, Objetivos e Hipóteses

Pesquisas recentes na área de SAT de artigos de notícias têm investigado diversas abordagens para essa tarefa adotando diferentes formas de modelar o problema. Os primeiros trabalhos focavam na sumarização monodocumento. Contudo, após o crescimento de informações disponíveis e o desenvolvimento da *Web* 2.0, o foco foi sendo alterado para a tarefa de sumarização multidocumento. Apesar dessa mudança, a sumarização monodocumento ainda é um problema em aberto (TORRES-MORENO, 2014) e muitas das recentes técnicas propostas para o contexto multidocumento têm sido pouco exploradas para a SAT monodocumento.

Por ser uma área de pesquisa crescente e com uma intensa comunidade, existem diversos trabalhos que exploraram vários tipos de métodos de sumarização. Dentre esses trabalhos, as abordagens baseadas em conceitos utilizando PLI (GILLICK et al., 2009; LI; QIAN; LIU, 2013; BOUDIN; MOUGARD; FAVRE, 2015; CAO et al., 2015; LI; LIU; ZHAO, 2015) têm ganhado destaque nos últimos anos, por apresentarem um bom desempenho, particularmente para a sumarização multidocumento. Esse tipo de abordagem trata o processo de sumarização como um problema de máxima cobertura, cujo objetivo é selecionar o subconjunto de sentenças do(s) documento(s) de entrada que maximiza a cobertura de conceitos relevantes, respeitando o tamanho máximo do resumo desejado. A maioria dos trabalhos existentes adota unidades textuais, como unigramas e bigramas, para representar a noção de conceitos. Somente em (SCHLUTER; SØGAARD, 2015) é possível observar uma comparação entre diversas formas de representação no contexto da sumarização multidocumento. Diferentes métodos individuais e combinados têm sido propostos para mensurar a relevância dos conceitos extraídos. Contudo, os trabalhos existentes na literatura ou utilizam um único método individual (GILLICK et al., 2009; BOUDIN; MOUGARD; FAVRE, 2015) ou exploram a combinação de diferentes métodos de forma supervisionada (LI; QIAN; LIU, 2013; CAO et al., 2015; LI; LIU; ZHAO, 2015), aplicando algoritmos de regressão

para estimar a relevância dos conceitos. Dessa forma, a combinação não supervisionada de diferentes métodos de ponderação de conceitos ainda é pouco explorada neste cenário. Embora existam alguns trabalhos que utilizam PLI para a sumarização monodocumento (HIRAO et al., 2013; PARVEEN; RAMSL; STRUBE, 2015a; PARVEEN; STRUBE, 2015b), no melhor do conhecimento do autor deste trabalho, nenhuma dessas abordagens adotou uma modelagem baseada em conceitos, conforme proposto por Gillick et al. (2009).

Apesar dos avanços obtidos nos últimos anos, principalmente na tarefa de sumarização multidocumento, ainda existe uma grande disparidade entre a qualidade dos resumos gerados automaticamente e os escritos manualmente por seres humanos. Rankel et al. (2011) mencionam que os avaliadores da competição de sumarização *Text Analysis Conference* (TAC), do ano de 2008, confidenciaram que facilmente distinguiam os resumos gerados automaticamente, mesmo sem terem lido os resumos de referência. Os avaliadores humanos afirmaram que os resumos produzidos automaticamente falhavam nos dois aspectos essências de um resumo: *Informatividade* e *Qualidade Linguística*. A informatividade reflete a quantidade de informações relevantes do(s) documento(s) de entrada que estão presentes no resumo gerado. Enquanto isso, a qualidade linguística mensura os aspectos da legibilidade do resumo, como coesão, coerência e gramaticalidade.

Atualmente, a maioria das abordagens de SAT extrativas focam somente em maximizar a informatividade dos resumos gerados, enquanto que poucos trabalhos tratam também a qualidade linguística dos resumos produzidos (CHRISTENSEN et al., 2013; PARVEEN; STRUBE, 2015b). Embora a seleção de conteúdo relevante para compor o resumo de saída seja o aspecto mais pesquisado, ele ainda permanece um problema em aberto. Como afirmado pelos avaliadores do TAC 2008, os resumos escritos manualmente possuíam muito mais informações relevantes do que os melhores resumos gerados automaticamente (RANKEL et al., 2011). Esse fato também pode ser observado analisando os resumos criados por seres humanos e pelos mais recentes sistemas de SAT extrativos do estado da arte, em ambas as tarefas de sumarização monodocumento e multidocumento.

A análise da literatura conduzida neste trabalho evidenciou que em sua grande maioria, as atuais abordagens SAT extrativas de artigos de notícias são estáticas, ou seja, adotam um método de sumarização com um conjunto de parâmetros pré-definido para todos os documentos de entrada. Essa característica é uma significante limitação das atuais abordagens de SAT, já que nenhum método de sumarização consegue produzir resumos informativos para todos os documentos de entrada, mesmo quando eles pertencem ao mesmo domínio (HONG; MARCUS; NENKOVA, 2015).

O cenário exposto motivou o desenvolvimento deste trabalho de doutorado, cujo foco principal é investigar como aumentar a informatividade dos resumos gerados automaticamente. Visando suprir as limitações supracitadas, o objetivo geral deste trabalho é propor e avaliar uma abordagem baseada em conceitos utilizando PLI e regressão para as tarefas de sumarização monodocumento e multidocumento de artigos de notícias. A

arquitetura da solução proposta é dividida em duas etapas principais: **(i)** a geração de diversos resumos candidatos e, a seguir **(ii)** a predição e seleção do resumo mais informativo. Duas abordagens baseadas em conceitos usando PLI são propostas para a geração dos resumos candidatos para as tarefas de sumarização monodocumento e multidocumento. As abordagens desenvolvidas possibilitam a exploração de diversas configurações, permitindo a geração de uma grande variedade de resumos candidatos representando diferentes perspectivas. Tais abordagens são integradas em uma macro arquitetura com a fase de identificação e seleção do resumo mais informativo. Essa etapa de seleção visa estimar a cobertura de informações relevantes nos resumos candidatos gerados, possibilitando, assim, a identificação do resumo estimado como mais representativo. Para isso, diversas características baseadas em tradicionais indicadores de relevância, como posição, frequência e centralidade, identificadas na literatura, e outras introduzidas neste trabalho são exploradas para a construção de um modelo de regressão.

Para alcançar o objetivo geral definido neste trabalho, alguns objetivos específicos são elencados a seguir.

1. Conduzir uma análise empírica de diversas técnicas e estratégias de combinações para a tarefa de pontuação da relevância das sentenças, buscando verificar quais são as melhores técnicas para cada uma das tarefas de sumarização monodocumento e multidocumento.

2. Propor uma abordagem baseada em conceitos utilizando PLI para a tarefa de sumarização monodocumento que combina os aspectos de posição das sentenças e centralidade em nível de sentenças para ponderar a importância dos conceitos extraídos.

3. Apresentar uma abordagem baseada em conceitos utilizando PLI para a tarefa de sumarização multidocumento, que explora os aspectos de posição das sentenças e centralidade em nível de documento para mensurar a relevância dos conceitos extraídos e filtrar sentenças com baixo grau de centralidade.

4. Investigar e definir características que possam ser adotadas para estimar, em nível de resumo, a quantidade de informações relevantes presentes em um resumo.

5. Propor uma macro abordagem que integra as metodologias de sumarização monodocumento e multidocumento propostas nos objetivos **(2)** e **(3)**, respectivamente, em uma única arquitetura baseada em duas etapas. A primeira etapa utiliza as abordagens baseadas em conceitos usando PLI para a geração de diversos resumos candidatos, dependendo da tarefa de sumarização em questão. Posteriormente, em uma segunda etapa, cada resumo candidato extraído é analisado, visando estimar o seu grau de informatividade, aplicando o modelo de regressão treinado usando as características definidas no objetivo **(4)**.

Este trabalho é fundamentado na hipótese principal de que é possível aumentar a informatividade dos resumos gerados automaticamente, nas tarefas de sumarização monodocumento e multidocumento, adotando uma estratégia de sumarização que produz diversos resumos candidatos explorando diferentes configurações em uma abordagem baseada em conceitos usando PLI e, posteriormente, analisa e seleciona o resumo candidato estimado como mais informativo aplicando algoritmos de regressão. Além dessa, outras hipóteses secundárias também são investigadas neste trabalho, sendo elas:

1. A combinação dos métodos de posição e frequência das sentenças para mensurar a importância dos conceitos, em uma abordagem usando PLI na tarefa de sumarização monodocumento, aumenta a informatividade dos resumos gerados.

2. O uso de uma estratégia de distribuição de pesos que somente pontua a primeira ocorrência dos conceitos extraídos em uma abordagem utilizando PLI aumenta a informatividade dos resumos produzidos na sumarização monodocumento.

3. A integração dos métodos de posição das frases e frequência dos documentos para ponderar os conceitos, em uma abordagem usando PLI na tarefa de sumarização multidocumento, produz resumos mais informativos.

4. A filtragem de sentenças com um baixo grau de centralidade melhora o desempenho das abordagens baseadas em conceitos utilizando PLI, em termos de tempo de execução e informatividade dos resumos produzidos.

5. O uso individual e combinado dos métodos de ponderação de conceitos (posição das frases, frequência dos documentos e frequência das sentenças), em conjunto com medidas de similaridade e divergência, melhora o desempenho do processo de estimação da informatividade de um resumo.

Para o desenvolvimento deste trabalho, utilizou-se a combinação da modelagem baseada em conceitos usando PLI proposta por Gillick et al. (2009), para a geração de diversos resumos candidatos, em conjunto com a estratégia de estimar a informatividade de cada candidato gerado utilizando algoritmos de regressão. A motivação para adotar a modelagem baseada em conceitos usando PLI vem dos bons resultados obtidos por diversos sistemas do estado da arte (GILLICK et al., 2009; LI; LIU; ZHAO, 2015; BOUDIN; MOUGARD; FAVRE, 2015; CAO et al., 2015) na tarefa de sumarização multidocumento. Além disso, esse tipo de abordagem permite gerar uma grande variedade de resumos, explorando diferentes configurações, como a forma de representação e o método de ponderação de conceitos.

A motivação para adotar algoritmos de regressão para estimar a informatividade de um resumo vem do bom desempenho obtido na tarefa de sumarização multidocumento reportado em (HONG; MARCUS; NENKOVA, 2015). Nesse trabalho, os autores apresentaram uma abordagem que primeiro produz diversos resumos candidatos, combinando em

nível de sentença diversos resumos gerados por quatro sistemas de sumarização do estado da arte. Em uma segunda etapa, um algoritmo de regressão é aplicado para estimar a cobertura de informações relevantes em cada resumo candidato gerado. Além disso, em uma linha de pesquisa diferente, Louis e Nenkova (2009), Saggion et al. (2010) investigaram a estratégia de avaliar sistemas de sumarização sem adotar resumos de referência. Para isso, os autores exploraram diversas características extraídas dos resumos gerados e das relações de similaridade e divergência com os documentos originais. Ambos os trabalhos obtiveram resultados encorajadores, demonstrando que existia uma forte correlação das características investigadas com os sistemas PYRAMID (NENKOVA; PASSONNEAU; MCKEOWN, 2007) e *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) (LIN, 2004), que são tradicionalmente adotados para avaliar sistemas de sumarização.

Apesar de focar exclusivamente na tarefa de sumarização de artigos de notícias, os métodos investigados e propostos neste trabalho também podem ser adaptados/estendidos para outros domínios, como artigos científicos, textos de blogs, entre outros domínios. Contudo, conforme apontado por Ferreira et al. (2013), os métodos de sumarização apresentam uma grande variação nos seus desempenhos, dependendo do tipo de documento a ser sumarizado. Dessa forma, melhores resultados podem ser obtidos explorando as características do(s) documento(s) de entrada para tomar a decisão de quais técnicas de sumarização adotar.

1.3 Contribuições do Trabalho

As principais contribuições deste trabalho são resumidas a seguir:

- **Contribuição 1:** Investigação do desempenho de diversas técnicas superficiais para computar a importância das sentenças e estratégias de combinação, considerando as tarefas de sumarização monodocumento e multidocumento.
- **Contribuição 2:** Realização de uma análise comparativa no desempenho de diversas formas de representação e métodos de ponderação de conceitos em uma abordagem utilizando PLI, nas tarefas de sumarização monodocumento e multidocumento.
- **Contribuição 3:** Criação de uma abordagem baseada em conceitos usando PLI para a sumarização monodocumento. A solução proposta adota um novo método para a ponderação da relevância dos conceitos (bigramas) que combina os aspectos de posição e frequência das sentenças, além de considerar a importância dos conceitos adjacentes durante o processo. Uma nova estratégia para a distribuição dos pesos dos conceitos que pontua somente a primeira ocorrência também é apresentada. Além disso, a abordagem proposta adota o modelo de Grafo de Entidades (GUINAUDEAU; STRUBE, 2013) para selecionar sentenças que maximizam uma aproximação da coesão local do resumo gerado.

- **Contribuição 4:** Desenvolvimento de uma abordagem baseada em conceitos usando PLI para a sumarização multidocumento, que considera os aspectos de centralidade e posição para ponderar a relevância dos conceitos (bigramas) extraídos e filtrar sentenças não relevantes. A solução proposta aplica um novo método de ponderação de conceitos que combina os aspectos de centralidade e a posição, e também explora os títulos dos documentos para melhor estimar a relevância dos conceitos extraídos. Além disso, uma estratégia de filtragem baseada no percentual de documentos de entrada que remove grupos de sentenças com poucos membros é adotada.

- **Contribuição 5:** Concepção de uma abordagem baseada em conceitos utilizando PLI e regressão para as tarefas de sumarização monodocumento e multidocumento. A arquitetura da abordagem proposta é dividida em duas etapas centrais: **(i)** Geração dos resumos candidatos; e **(ii)** Seleção do resumo mais informativo. A geração dos resumos candidatos é executada explorando diversas configurações (formas de representação e métodos de ponderação) em conjunto com as abordagens baseadas em conceitos utilizando PLI definidas anteriormente, dependendo da tarefa de sumarização a ser executada. Em seguida, um algoritmo de regressão é aplicado para estimar o resumo que possui a maior cobertura de informações relevantes, visando selecionar o resumo estimado como mais informativo. Para treinar o modelo de regressão proposto, inicialmente, definiu-se um conjunto de características com base em indicadores individuais e combinados de importância de conteúdo, como posição, frequência e centralidade, em o conjunto com diversas medidas de similaridade e divergência computados entre o resumo e o(s) documento(s) de entrada.

1.4 Estrutura do Documento

Este trabalho de doutorado está organizado em sete capítulos. Uma visão geral da estrutura do documento e das dependências entre seus capítulos é apresentada na Figura 1. Os capítulos e o apêndice dentro da caixa tracejada no centro da Figura 1 apresentam as contribuições desenvolvidas neste trabalho. As linhas tracejadas indicam uma dependência entre os elementos do documento, ou seja, um capítulo ou apêndice deve ser lido antes do outro para uma melhor compreensão. Por exemplo, o Capítulo 4 e o Capítulo 5 devem ser lidos somente depois da leitura do Capítulo 3. O Apêndice A só deve ser lido após o Capítulo 4. Por fim, o Capítulo 6 só deve ser lido após a leitura do Capítulo 4, Capítulo 5 e do Apêndice A.

Uma breve descrição dos capítulos e do apêndice deste trabalho de doutorado é apresentado a seguir:

- **Capítulo 2 - Sumarização Automática de Textos:** Esse capítulo apresenta os conceitos fundamentais para o entendimento da área de SAT, sendo estruturado em

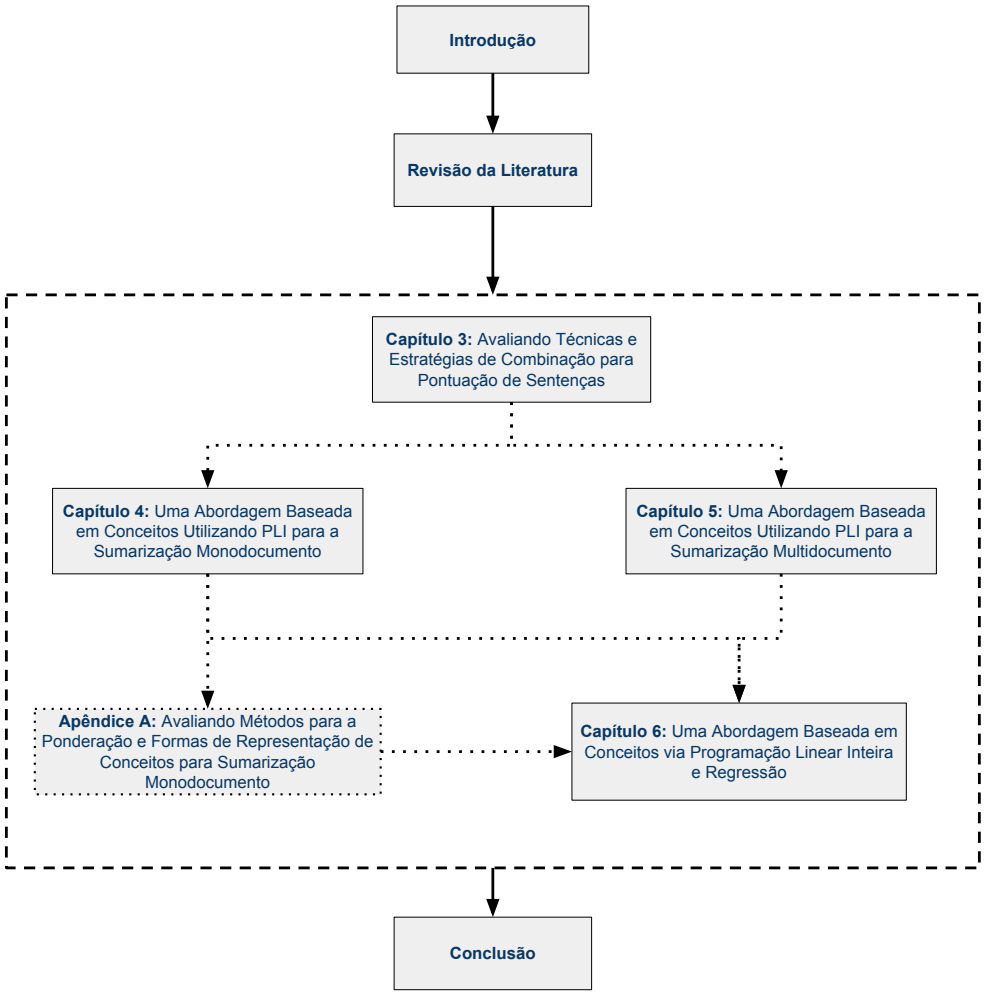


Figura 1 – Visão geral da estrutura do documento. Os capítulos são apresentados por caixas conectadas. As setas sólidas indicam a sequência dos capítulos que podem ser lidos sem nenhuma dependência com outro. As setas tracejadas indicam os elementos (capítulo ou apêndice) que devem ser lidos antes do elemento apontado para uma melhor compreensão.

seis seções. A primeira seção apresenta uma breve contextualização histórica da área de SAT. Na segunda seção serão introduzidas diversas dimensões comumente usadas para caracterização da área de SAT. A terceira seção discutirá os principais trabalhos relacionados identificados para a sumarização monodocumento e multidocumento de artigos de notícias. Os principais métodos adotados para a avaliação de sistemas de SAT serão apresentados na quarta seção. Por fim, as principais medidas de avaliação de resumos serão apresentadas na última seção.

- **Capítulo 3 - Avaliando Técnicas e Estratégias de Combinação para a Pontuação de Sentenças:** Neste capítulo são apresentados os experimentos realizados para avaliar diferentes técnicas superficiais para a pontuação de sentenças e estratégias de combinação dessas para a sumarização monodocumento e multidocumento. Dezoito técnicas de pontuação de sentenças serão investigadas, em vários cenários

de aplicação: individualmente, aplicando estratégias para combiná-las e adotando algoritmos de Aprendizagem de Máquina.

- **Capítulo 4 - Uma Abordagem Baseada em Conceitos Utilizando PLI para a Sumarização Monodocumento:** Esse capítulo apresenta a proposta de uma abordagem baseada na maximização de conceitos relevantes utilizando PLI para a sumarização monodocumento. A solução proposta combina os métodos de posição e frequência das sentenças para ponderar a relevância (pesos) dos conceitos (bigramas) extraídos. Além disso, será apresentada uma estratégia de distribuição de pesos que atribui o escore de importância somente para a primeira ocorrência de um conceito e zero nas demais ocorrência. O modelo de Grafo de Entidade (GUINAUDEAU; STRUBE, 2013) também será aplicado para gerar uma aproximação da coesão local do resumo gerado.

- **Capítulo 5 - Uma Abordagem Baseada em Conceitos Utilizando PLI para a Sumarização Multidocumento:** Neste capítulo é apresentada a proposta da abordagem baseada em conceitos utilizando PLI para a tarefa de sumarização multidocumento. A abordagem desenvolvida combina os métodos de posição das sentenças e frequência dos documentos para mensurar a importância dos conceitos (bigramas) extraídos. Além disso, será apresentada a estratégia baseada em centralidade criada para filtrar sentenças com baixo escore de centralidade, visando reduzir o tempo de execução do processo de sumarização e também aumentar a informatividade do resumo gerado.

- **Capítulo 6 - Uma Abordagem Baseada em Conceitos via PLI e Regressão:** Neste capítulo é apresentada a proposta da abordagem supervisionada que combina PLI e regressão para as tarefas de sumarização monodocumento e multidocumento. A solução proposta, inicialmente, produz diversos resumos candidatos adotando as abordagens baseadas em conceitos propostas no Capítulo 4 e no Capítulo 5, dependendo da tarefa de sumarização a ser executada. Em uma segunda etapa, um algoritmo de regressão é aplicado para estimar a quantidade de informações relevantes presentes nos resumos candidatos gerados, visando identificar o resumo estimado como mais informativo.

- **Capítulo 7 - Conclusão:** Por fim, neste capítulo são apresentadas as conclusões obtidas a partir das investigações realizadas, as contribuições do trabalho desenvolvido, as limitações observadas, e são delineadas algumas linhas de trabalhos futuros que podem ser seguidas.

- **Apêndice A - Avaliando Métodos de Ponderação e Representação de Conceitos para a Sumarização Monodocumento:** Nesse Apêndice são apresentados

os experimentos realizados para avaliar diferentes formas de representação e métodos de ponderação da importância dos conceitos em uma abordagem usando PLI para a sumarização monodocumento.

2 SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS

Neste capítulo, é apresentada uma breve introdução dos principais conceitos que envolvem a área de Sumarização Automática de Textos (SAT). Inicialmente, uma contextualização histórica do surgimento e uma caracterização da área são introduzidas na Seção 2.1 e Seção 2.2, respectivamente. Em seguida, as principais abordagens identificadas na literatura para a sumarização monodocumento e multidocumento são descritas na Seção 2.3. Na Seção 2.4 são apresentados os principais métodos adotados para a avaliação de sistemas de SAT. Por fim, na Seção 2.5 os principais desafios da área são realçados e quais desses são abordados nesta tese.

2.1 Contextualização Histórica

O ato de sintetizar (sumarizar) algo é um processo comum no cotidiano dos seres humanos. Um exemplo clássico acontece quando se conversa com alguém e esta pessoa pergunta como foi o seu dia ou o que aconteceu na última reunião do grupo de pesquisa. Ao invés de contar todos os mínimos detalhes, geralmente, faz-se uma síntese dos principais acontecimentos. Outro exemplo prático são os resumos de documentos escritos, presentes em artigos científicos ou textos de notícias. Esses resumos são particularmente úteis, pois, auxiliam o leitor que primeiro lê o resumo para então decidir se o documento é relevante ou não para os seus interesses.

A sumarização realizada pelos seres humanos é um processo cognitivo e requer um completo entendimento do documento ou fato a ser sintetizado (TORRES-MORENO, 2014). Esse processo não é exato, ou seja, se diferentes pessoas sumarizarem um mesmo documento, é provável que resumos diversos sejam gerados. Tal comportamento demonstra o quão complexa e abstrata é a tarefa de sumarização, e quão desafiadora é a ideia de automatizá-la.

A Sumarização Automática de Textos (SAT) pode ser definida como o processo automático que visa gerar uma versão mais compacta (resumo) de um ou mais textos fonte, mantendo suas informações mais relevantes (NENKOVA; MCKEOWN, 2012). A primeira iniciativa na área SAT data de 1958 com o trabalho pioneiro de *Hans Peter Luhn*, na tentativa de sumarização automática de artigos científicos (LUHN, 1958). Em seu trabalho, Luhn (1958) apontava a importância que as palavras mais frequentes do documento de entrada possuem para identificar informações importantes que deveriam estar nos resumos gerados. Edmundson (1969) expandiu o trabalho de Luhn propondo três novos métodos para identificar sentenças importantes para compor o resumo de um documento. Em seu

trabalho, Edmundson propôs uma combinação linear ponderada de quatro métodos **(i)** frequência das palavras, **(ii)** sobreposição das palavras de cada sentença com o título do documento, **(iii)** posição das sentenças no documento, e **(iv)** presença de expressões chave na sentença. O trabalho de Edmundson é considerado como a primeira tentativa de combinar diferentes métodos para computar a importância de uma sentença para a SAT. Outros importantes trabalhos foram desenvolvidos por Rath, Resnick e Savage (1961), Edmundson e Wyllys (1961) e em anos posteriores por Earl (1970), Rush, Salvador e Zamora (1971), Paice (1981), Rau, Jacobs e Zernik (1989), entre outros.

Nos anos noventa, surgiram diversos trabalhos (ENDRES-NIGGEMEYER, 1990; ENDRES-NIGGEMEYER; HOBBS; JONES, 1993; ENDRES-NIGGEMEYER, 1993; JONES, 1993; KUPIEC; PEDERSEN; CHEN, 1995) que iniciaram um novo e crescente interesse na área. O crescimento da área de SAT no final dos anos noventa e início dos anos dois mil ocorreu devido ao desenvolvimento e a expansão de novas tecnologias, como a *Internet*, e também graças às diversas campanhas e competições criadas para estimular novas pesquisas e o desenvolvimento de sistemas para a SAT (SAGGION; POIBEAU, 2013). As iniciativas da TIPSTER em 1997, SUMMAC (MANI et al., 2002) em 1998 e *NII Testbeds and Community for Information access Research* (NTCIR) em 2001 promoveram o desenvolvimento e a avaliação de diversos sistemas de sumarização através da criação e disponibilização de conjunto de dados para validar o desempenho de sistemas de sumarização. Em 2005, a Multilingual SUmmarization Evaluation (MS) conduziu uma competição internacional para avaliar sistemas em Inglês e Árabe. As conferências do *Document Understanding Competition* (DUC) (OVER; DANG; HARMAN, 2007) dos anos de 2001 até 2007 promovidas pela *National Institute of Standards and Technology* (NIST), foram responsáveis pela criação e disponibilização de diversos outros corpora de artigos de notícias para a avaliação de sistemas de sumarização, que ainda hoje são muito utilizados (GAMBHIR; GUPTA, 2016). Mais recentemente, outros eventos como a *Text Analysis Conference* (TAC) (OWCZARZAK; DANG, 2011) e a MultiLing 2017 (GIANNAKOPOULOS et al., 2017) continuaram a promover competições que estimulam o desenvolvimento da área.

Atualmente, a SAT é uma intensa área de pesquisa, em particular para a comunidade de Processamento de Linguagem Natural (PLN) e Mineração de Texto (MT). A literatura da área de SAT é vasta e diversificada, sendo possível identificar pesquisas envolvendo diversos métodos desenvolvidos por outras áreas, como Recuperação de Informação (RI), Extração de Informação (EI), Aprendizagem de Máquina (AM), Análise de Discurso, Geração de Linguagem Natural (GLN), além de outras áreas relacionadas.

Na seção a seguir são apresentadas diversas dimensões que são comumente usadas para caracterização da área de SAT.

2.2 Caracterização da Área

O termo sumarização, de maneira geral, se refere à tarefa de gerar um resumo contendo as informações mais relevantes, a partir de um conjunto de documentos de entrada. Dessa forma, o processo de sumarização envolve uma grande variedade de tarefas que podem ser caracterizadas com base em diversas dimensões, tais como a função do resumo gerado, a quantidade e o tipo dos documentos de entrada, a tarefa de sumarização a ser realizada, o tipo da abordagem de sumarização adotada, entre outras (LLORET; PALOMAR, 2012).

Além da sumarização de documentos textuais, existem trabalhos que processam outros tipos de documentos, como imagens (CAMARGO; GONZáLEZ, 2016), vídeos (GUO et al., 2016) e voz (KOZIEL et al., 2015). Com relação à quantidade de documentos de entrada, as abordagens podem ser classificadas em **Monodocumento** e **Multidocumento**. Na sumarização monodocumento, um resumo deve ser criado a partir de um único documento de entrada, enquanto que na sumarização multidocumento um único resumo deve ser gerado analisando vários documentos de entrada relacionados com um mesmo evento ou assunto. Cada uma dessas tarefas envolve desafios distintos, por exemplo, para a sumarização multidocumento a redundância de informações entre os documentos é muito alta, sendo essencial a adoção de estratégias para evitar a presença de informações redundantes no resumo gerado.

Os resumos gerados podem ser classificados de acordo com sua funcionalidade, como **Indicativos**, **Informativos**, ou **Críticos** (SAGGION; POIBEAU, 2013; GAMBHIR; GUPTA, 2016). Os resumos indicativos contêm apenas a indicação dos principais tópicos discutidos no(s) documento(s). Geralmente esse tipo de resumo é parecido como uma tabela de conteúdo, listando os tópicos mais relevantes. Enquanto isso, os resumos informativos têm por objetivo apresentar as informações mais relevantes do(s) documento(s) de entrada, apresentando os tópicos mais importantes com um pouco mais de informações. A ideia desse tipo de resumo é ser uma versão mais compacta do documento de entrada, e, em alguns casos, eles podem substituir a leitura dos documentos originais. Os resumos críticos apresentam, de forma sucinta, avaliações, comparações e opiniões sobre produtos, serviços, ou outros tópicos dos documentos originais. Esse tipo de resumo também possui associada a polaridade (positiva, negativa ou neutra) das avaliações de cada assunto discutido nos textos de entrada.

Com base no idioma, existem três tipos de resumos (LLORET; PALOMAR, 2012): **monolíngue**, **multilíngue** e **idiomas cruzados**. A sumarização mais comum é a monolíngue, que ocorre quando o idioma do resumo a ser gerado é o mesmo do(s) documento(s) de entrada. A sumarização multilíngue ocorre quando o(s) documento(s) de entrada estão em um conjunto de idiomas qualquer, por exemplo, inglês, português ou francês, e o resumo também pode ser gerado em qualquer um desses idiomas. A sumarização de idiomas cruzados ocorre quando o conjunto de documentos de entrada está em um único idioma

e os resumos podem ser gerados em vários idiomas.

As abordagens também podem ser classificadas de acordo com o gênero dos documentos de entrada (LLORET; PALOMAR, 2012): (i) **Sumarização de Notícias**: resumos de artigos de notícias; (ii) **Sumarização Especializada**: resumos de documentos especializados em um único domínio, por exemplo, Ciência, Medicina, Biologia, entre outros; (iii) **Sumarização Literária**: resumos de documentos narrativos, textos literários, entre outros; (iv) **Sumarização de Enciclopédias**: sumários de documentos de enciclopédias, por exemplo, a Wikipédia; (v) **Sumarização em Redes Sociais**: resumos de postagens em redes sociais; (vi) **Sumarização de atas de reuniões**: resumos de textos produzidos em uma ou mais reuniões (BANERJEE; MITRA; SUGIYAMA, 2015a); entre outros tipos de documentos.

As abordagens também podem ser classificadas de acordo com o seu propósito (SAGGION; POIBEAU, 2013): (i) **Sumarização Genérica**: o processo de sumarização é executado sem levar em consideração quaisquer informações que o usuário necessita; (ii) **Sumarização Baseada em Consultas**: o processo de sumarização é realizado considerando as informações necessárias para responder uma determinada consulta do usuário; (iii) **Sumarização de Atualizações**: esse caso assume que o leitor já possui informações prévias sobre o tópico dos documentos, e eles desejam apenas atualizações das informações existentes; e (iv) **Sumarização baseada em sentimentos**: essa tarefa é a união da SAT com a área de Análise de Sentimentos (PANG; LEE, 2008). Os resumos gerados nesse tipo de sumarização envolvem o processo classificação dos documentos quanto a sua subjetividade, geração dos resumos, e classificação quanto a polaridade das informações como positivas, negativas ou neutras.

Além dos resumos, as abordagens de SAT também podem ser classificadas como **Extrativas** ou **Abstrativas** (NENKOVA; MCKEOWN, 2012). Abordagens extrativas selecionam as sentenças mais relevantes dos documentos de entrada e as utilizam sem nenhuma alteração para compor o resumo. Por outro lado, as abordagens Abstrativas focam na identificação e seleção das informações mais importantes dos documentos e as reescrevem de forma mais sucinta utilizando operações como compressão de sentenças (ZAJIC et al., 2007), fusão de sentenças (FILIPPOVA, 2010) e geração de linguagem natural (KHAN; SALIM; KUMAR, 2015).

O Quadro 1 apresenta um resumo das dimensões supracitadas e as possíveis classificações.

Quadro 1 – Caracterização das abordagens de SAT baseada em diversas características.

Quantidade de documentos	Monodocumento Multidocumento
Tipo de Abordagem	Extrativa Abstrativa
Tipo de Entrada	Textos Imagens Vídeos Voz Hipertexto
Tipo de Resumo	Indicativo Informativo Crítico
Propósito	Genérica Atualização Baseada em Consultas Baseada em Sentimentos
Idioma	Monolíngue Multilíngue Idiomas cruzados

2.3 Principais Trabalhos Relacionados

Nesta seção, são apresentados os principais trabalhos relacionados no contexto das tarefas de sumarização genérica monodocumento e multidocumento de artigos de notícias escritos em Inglês. Na Subseção 2.3.1 são apresentados os trabalhos na tarefa de sumarização monodocumento e na Subseção 2.3.2 os trabalhos desenvolvidos no contexto da sumarização multidocumento.

2.3.1 Sumarização Monodocumento

Os primeiros trabalhos na área de SAT focavam na criação automática de resumos a partir de um único documento de entrada (TORRES-MORENO, 2014). Nas conferências do DUC, a sumarização monodocumento foi o foco das competições dos anos de 2001 e 2002. Nessas competições, nenhum sistema automático conseguiu obter resultados estatisticamente superiores ao *baseline*, que consistia em usar as 100 primeiras palavras do texto como resumo (NENKOVA, 2005). Nos anos seguintes, a tarefa foi praticamente descontinuada, dando lugar para a sumarização multidocumento. Apesar de atualmente a maioria das pesquisas na área de SAT serem focadas na sumarização multidocumento, os resultados obtidos pelos atuais sistemas de sumarização monodocumento indicam que ela

ainda é um problema em aberto (TORRES-MORENO, 2014).

Diversos trabalhos investigaram a aplicação de técnicas superficiais para a pontuação das sentenças para a SAT. Essas técnicas se destacaram porque são simples de serem desenvolvidas, exigem pouco ou nenhum recurso linguístico e possuem um baixo processamento computacional. Algumas das principais técnicas superficiais adotadas na literatura incluem frequência das palavras, posição das sentenças, similaridade das sentenças com o título do documento, entre várias outras (FERREIRA et al., 2013). O primeiro trabalho nessa área, desenvolvido por Luhn (1958), era baseado no pressuposto de que as palavras mais frequentes do documento eram bons indicativos do principal tema discutido. Para isso, palavras muito frequentes (*stopwords*), como artigos, preposições ou pronomes, eram removidas. Posteriormente, computava-se a frequência das palavras restantes e as sentenças com o maior número de palavras frequentes recebiam maior probabilidade de serem incluídas nos resumos. Ferreira et al. (2014) investigaram várias combinações de técnicas de pontuação de sentença superficiais em três domínios: textos de blogs, documentos de notícias e artigos científicos. Os autores analisaram e identificaram as melhores combinações para cada domínio investigado. No domínio de artigos de notícias, o melhor desempenho foi obtido por meio da combinação dos métodos de Frequência dos Termos e Frequência Inversa dos Termos (TF-IDF), similaridade léxica entre as sentenças, posição das sentenças no documento e a similaridade das sentenças com o título.

Outros trabalhos exploraram a combinação ponderada de diversas técnicas de pontuação de sentenças e utilizaram algoritmos evolucionários para identificar as melhores configurações de pesos para cada método adotado. García-Hernández e Ledeneva (2013) investigaram a combinação dos métodos de posição das sentenças e frequência das palavras para a pontuação das sentenças. Para otimizar os resultados, algoritmos genéticos foram utilizados para estimar a melhor configuração de pesos para cada uma das técnicas adotadas. Fattah e Ren (2009) propuseram uma abordagem para a sumarização monodocumento utilizando diversas técnicas de pontuação de sentenças, como frequência de palavras chave positivas e negativas, posição das sentenças, centralidade das frases, entre outras. Os autores investigaram as técnicas de pontuação individualmente, e, posteriormente, exploraram diversas combinações ponderadas dessas técnicas. Para a obtenção de uma configuração adequada dos pesos, algoritmos genéticos e modelos de regressão foram aplicados. As técnicas de pontuação investigadas foram utilizadas como características de treinamento para a construção de um modelo de sumarização utilizando redes neurais sem realimentação (*Feedforward neural network*), redes neurais probabilísticas (*Probabilistic neural network*) e modelos de misturas gaussianas (*Gaussian mixture model*).

Mendoza et al. (2014) apresentaram um método para a sumarização monodocumento usando operadores genéticos e pesquisa local guiada. O método proposto utiliza um algoritmo memético que combina a estratégia de pesquisa baseada na população gerada pelo algoritmo evolucionário com uma estratégia de busca local guiada. O processo de suma-

rização foi tratado como um problema de otimização binária, na qual a informatividade dos resumos gerados é estimada com base na ponderação de diversas características de cada sentença do documento, como posição, tamanho, similaridade com o título e com um grupo de características envolvendo a centralidade das sentenças.

Algoritmos baseados em grafos também são muito explorados para a sumarização monocumento (FERREIRA et al., 2013; GAMBHIR; GUPTA, 2016). Esse tipo de abordagem explora as relações entre os elementos do documento, como sentenças e palavras, para identificar as informações mais importantes para compor o resumo gerado. Alguns dos principais algoritmos adotados são o LexRank (ERKAN; RADEV, 2004) e o TextRank (MIHALCEA; TARAU, 2004). O algoritmo *LexRank* computa a importância das sentenças baseada na centralidade dos autovetores (*Eigenvector Centrality*) em um grafo representando as sentenças do documento. Esse modelo utiliza uma matriz de semelhança baseada na similaridade do cosseno entre as sentenças. Se essa similaridade for maior que um determinado limiar, então uma aresta entre essas duas sentenças é criada no grafo. Por fim, as frases mais importantes são selecionadas pelo algoritmo através da realização de um caminhamento aleatório no grafo gerado. O *TextRank* é outro popular algoritmo baseado em grafos muito usado para a sumarização monodocumento e extração de palavras chave. Esse algoritmo é baseado no tradicional algoritmo *Pagerank* (BRIN; PAGE, 1998) e explora a ideia de representar as sentenças do documento como vértices em um grafo e arestas são criadas utilizando a similaridade do cosseno entre as duas sentenças. O TextRank define a importância de uma sentença (vértice) com base na relevância dos vértices vizinhos da sentença no grafo criado.

Trabalhos mais recentes têm explorado a ideia de modelar o processo de sumarização como um problema de otimização combinatória com restrições, ou seja, maximizar aspectos essenciais dos resumos, por exemplo, informatividade e coesão, atendendo ao mesmo tempo diversas restrições impostas, como o tamanho máximo do resumo a ser gerado. Hirao et al. (2013) apresentaram uma abordagem para a sumarização monodocumento que explora as dependências entre as sentenças extraídas a partir de estruturas retóricas obtidas usando o modelo da Teoria de Estrutura Retórica (*Rhetorical Structure Theory*) (MANN; THOMPSON, 1988). Além disso, dependências entre as palavras obtidas a partir da análise de dependência do documento também são utilizadas. Árvores aninhadas são criadas para representar as relações entre as sentenças e as palavras do documento. Por fim, os autores formularam a tarefa de sumarização como um problema de otimização com restrições, e utilizam Programação Linear Inteira (PLI) para selecionar as sentenças para compor o resumo.

Nessa mesma linha de pesquisa, Kikuchi et al. (2014) propuseram uma abordagem para sumarização monodocumento que faz uso das dependências entre as frases obtidas através de estruturas retóricas e dependências entre palavras geradas por meio de um analisador de dependência. Os dois tipos de dependências são representados através da

construção de uma árvore aninhada para o documento, que é composta de dois tipos de estruturas: **(i)** a árvore do documento, em que os vértices representam dependências entre as frases; e **(ii)** uma árvore da frase, na qual os vértices representam dependências entre as palavras. Uma árvore aninhada é construída substituindo os vértices na árvore de um documento por uma representando a frase. A tarefa de sumarização é então formulada como um problema de otimização, no qual a árvore aninhada deve ser filtrada sem perder o conteúdo importante no documento de origem.

Parveen, Ramsl e Strube (2015a) propuseram uma abordagem baseada em tópicos para sumarização monodocumento que maximiza conjuntamente os aspectos de informatividade, coesão e cobertura de tópicos. Os autores utilizaram um grafo bipartido em que as frases e tópicos são vértices, e as arestas entre eles são criadas para conectar uma frase com os tópicos que ela contém. Em seguida, o algoritmo *Hyperlink-Induced Topic Search* (HITS) é aplicado para ponderar essas frases. Finalmente, para estimar a coerência do resumo a ser gerado, os autores utilizaram um grafo de tópicos e aplicaram uma projeção ponderada sobre ele para identificar sobreposição de tópicos entre as sentenças. Parveen e Strube (2015b) modificaram o trabalho anterior, usando frases e entidades como vértices em um grafo bipartido em vez de tópicos. Os autores mudaram a representação baseada em tópicos por outra, utilizando o modelo de Grafo de Entidades (GUINAUDEAU; STRUBE, 2013) para obter uma estimativa mais aproximada da coesão local entre as frases. Em ambos os trabalhos, os autores avaliaram as abordagens propostas usando um corpus de artigos científicos no domínio de Medicina (*PLOS Medicine*) e o conjunto de dados do DUC 2002 composto por artigos de notícias.

Durrett, Berg-Kirkpatrick e Klein (2016) apresentaram uma abordagem para a sumarização monodocumento que combina compressão de sentenças e restrições que evitam problemas de correferências em aberto no resumo gerado. A abordagem proposta maximiza a seleção de unidades textuais no resumo gerado, ponderadas com base em um conjunto de atributos cujos pesos são aprendidos em um grande corpus de suporte. As unidades textuais adotadas são fragmentos das sentenças geradas usando regras de compressão e unidades de discurso com as respectivas *Rhetorical Structure Theory* (MANN; THOMPSON, 1987) relações entre elas. A abordagem criada aplica um conjunto de regras de compressão para excluir partes irrelevantes das sentenças originais do documento de entrada. Essas regras são desenvolvidas como dependências entre as unidades de textuais. Restrições de correferências em aberto são usadas para melhorar a coesão entre as sentenças, garantindo que para cada pronome incluído no resumo, a sentença contendo a entidade que ele menciona também esteja no resumo ou então o pronome é substituído pela entidade a qual ele se refere.

Lee e Lee (2017) apresentaram um modelo para a tarefa de sumarização monodocumento baseado em algoritmos de Aprendizagem por Reforço. Os autores exploram características embutidas (*embedding features*) relacionadas com a semântica distribucional

(*Word Embedding*) (PENNINGTON; SOCHER; MANNING, 2014), e a posição das frases do documento e aplicam um modelo baseado em Redes Q profundas (*Deep Q-Networks*) (MNIH et al., 2015) para realizar o processo de geração do resumo. No contexto da tarefa de sumarização, um estado denota um resumo que ainda pode estar incompleto (resumo parcial) e uma ação representa a inserção de uma nova sentença (sentença candidata) no resumo. O algoritmo de Redes Neurais Recursivas, do inglês *Recurrent Neural Network* (RNN), é usado para modelar o relacionamento entre um resumo parcial e uma sentença candidata a ser inserida. O processo de sumarização é realizado de forma gulosa inserindo a sentença candidata que maximiza um escore de relevância do resumo parcial, cujo valor é estimado aplicando uma função de regressão linear treinada adotando as características embutidas relacionadas às sentenças já presentes no resumo parcial.

No Quadro 2 são apresentados de forma resumida os principais trabalhos descritos anteriormente, sendo eles caracterizados utilizando as seguintes dimensões: **(i)** Tipo da abordagem de sumarização (Abstrativa e Extrativa); **(ii)** O processo de seleção das sentenças para compor o resumo é iterativo, guloso ou usa PLI; **(iii)** A abordagem desenvolvida é ou não supervisionada; e **(iv)** Principal vantagem e limitação do trabalho desenvolvido.

2.3.2 Sumarização Multidocumento

Com o crescente aumento na quantidade de documentos disponíveis, e também pelo fato de que, em muitos casos, vários deles estão discutindo o mesmo evento, a sumarização monodocumento perdeu popularidade, e pesquisadores mudaram o foco para a sumarização multidocumento nos últimos anos (SAGGION; POIBEAU, 2013). Um resumo gerado por um sistema de SAT multidocumento é uma breve representação das informações mais relevantes a partir de um conjunto de documentos relacionados. A relação entre os documentos pode ser de vários tipos, por exemplo, documentos podem ser relacionados porque eles discutem a mesma entidade, o mesmo tema, ou porque tratam o mesmo evento. Além do desafio de identificar as informações mais relevantes, outros dois problemas fundamentais que devem ser tratados por sistemas de sumarização multidocumento são a detecção e remoção de redundância e informações contraditórias (GAMBHIR; GUPTA, 2016).

Outro aspecto importante a ser tratado na sumarização multidocumento é a ordenação das sentenças (NENKOVA; MCKEOWN, 2012). Na sumarização monodocumento, presume-se que manter a ordem das informações no resumo assim como elas aparecem no documento entrada, em geral, produz bons resumos (SAGGION; POIBEAU, 2013). Em contrapartida, na sumarização multidocumento, uma atenção especial precisa ser dada para a ordem em que as sentenças serão inseridas no resumo, já que essas sentenças podem ser extraídas a partir de vários documentos distintos. Existem várias técnicas para lidar com a ordenação das frases, por exemplo, se a data da publicação do documento está disponível, então as sentenças podem ser apresentadas por ordem cronológica (GAMBHIR;

Quadro 2 – Resumo dos principais trabalhos relacionados para a tarefa de sumarização monodocumento.

Trabalho	**Abordagem**	**Método de Sumarização**	**Supervisionado**	**Vantagem**	**Limitação**
Fattah e Ren (2009)	Extrativa	Iterativa (sentenças)	Regressão e Classificação (sentenças)	Comparou diferentes modelos de classificação e regressão.	Nenhuma análise em nível de resumo é explorada. Avaliou considerando somente um corpus em Inglês e outro em Árabe.
Hirao et al. (2013)	Abstrativa	PLI (unidades discurso)	-	Adota técnicas de compressão de sentenças.	Considera somente frequência para mensurar relevância. Avaliou considerando somente um corpus.
García-Hernández e Ledeneva (2013)	Extrativa	Iterativa (sentenças)	Alg. Genéticos (sentenças)	Considera pesos diferentes para as técnicas de ponderação das sentenças.	Nenhuma análise em nível de resumo é explorada. Avaliou usando somente um corpus.
Ferreira et al. (2014)	Extrativa	Iterativa (sentenças)	-	Requer pouco processamento computacional.	Adotou uma combinação linear, sem pesos, fixa de técnicas de ponderação de sentenças. Adota um método de sumarização estático. Avaliou usando somente um corpus por domínio.
Kikuchi et al. (2014)	Abstrativa	PLI (sentenças, palavras)	-	Adota técnicas de compressão de sentenças.	Possui uma configuração estática. Avaliou usando somente um corpus.
Mendoza et al. (2014)	Extrativa	Iterativa (sentenças)	Alg. Genéticos (sentenças)	Considera pesos diferentes para as técnicas de ponderação das sentenças.	Inconsistência no desempenho da abordagem proposta nos dois corpora avaliados.

Parveen, Ramsl e Strube (2015a)	Extrativa	PLI (sentenças)	-	Trata os aspectos de informatividade, coesão e cobertura de tópicos.	Adota um método de sumarização estático. Avaliou usando somente um corpus por domínio.
Parveen e Strube (2015b)	Extrativa	PLI (sentenças)	-	Trata os aspectos de informatividade, coesão e não redundância.	Adota um método de sumarização estático. Avaliou usando somente um corpus por domínio.
Durrett, Berg-Kirkpatrick e Klein (2016)	Abstrativa	PLI (unidades de discurso)	Regressão (unidades de discurso)	Adota técnicas de compressão de sentenças. Trata os aspectos de informatividade e coesão.	Adota um método de sumarização estático. Avaliou usando somente um corpus.
Lee e Lee (2017)	Extrativa	Gulosa (sentenças)	Algoritmos de Regressão (sentenças)	Explora o uso de recurso relacionados a semântica distribucional (*Word Embedding*) e o relacionamento entre as sentenças para compor um resumo.	A estimação do escore de relevância do resumo considera apenas os aspectos de posição e semântica distribucional.

GUPTA, 2016). No entanto, isso nem sempre é possível, já que nem todos os documentos contêm sua data de publicação.

A ordenação das sentenças também pode ser realizada como uma forma de representar temas diferentes que devem estar contidos no resumo. Por exemplo, um algoritmo de agrupamento pode ser usado para identificar temas no conjunto de documentos de entrada e descobrir em que ordem os tópicos são apresentados nos documentos de entrada (BARZILAY; ELHADAD; MCKEOWN, 2002). Isso, por sua vez, pode ser utilizado para apresentar as frases em uma ordem semelhante à observada no conjunto de entrada. As abordagens probabilísticas procuram estimar a probabilidade de uma sequência de frases, tentando encontrar uma ordenação localmente ideal através da aprendizagem de restrições de ordenação de pares de sentenças (LAPATA, 2003).

Assim como na sumarização monodocumento, alguns dos primeiros trabalhos para a sumarização multidocumento focavam em métodos baseados na frequência das palavras para identificar informações relevantes. Nenkova, Vanderwende e McKeown (2006) propuseram o sistema ProbSum, que pondera a relevância das sentenças atribuindo a média das probabilidades das palavras dos documentos em relação às palavras da frase, com *stopwords* recebendo peso zero. Posteriormente, para evitar redundância, Hong et al. (2014) alteraram esse sistema adicionando a estratégia de somente incluir uma nova sentença no resumo caso ela não possua similaridade do cosseno maior que 0,5 com nenhuma outra sentença já presente no resumo. Outro trabalho nessa mesma linha de pesquisa foi o sistema LLRSum, proposto por Conroy, Schlesinger e O'Leary (2006). Esse sistema aplica o teste de proporção da verossimilhança, do inglês *Log-Likelihood Ratio* (LLR), para selecionar palavras chave (*topic words*) nos documentos de entrada. O teste LLR compara a distribuição de palavras dos documentos de entrada em relação a um outro *dataset*. O sistema então considera uma palavra como relevante se a sua estatística qui-quadrada obtida pelo teste LLR excede o valor 10. A importância de uma sentença é dada pelo número de palavras chave que ela possui dividido pelo seu total de palavras.

O sistema Greedy-KL (HAGHIGHI; VANDERWENDE, 2009) explora a suposição de que bons resumos possuem uma baixa divergência com o conjunto de documentos de entrada. Para isso, o objetivo desse sistema é encontrar o resumo que possui a menor pontuação utilizando a medida de divergência *Kullback-Leibler* (KL) entre a distribuição de probabilidades entre o resumo gerado e os documentos de entrada. Como otimizar globalmente o menor valor da medida KL é um problema de otimização com complexidade exponencial ($O(2^n)$) em relação ao total de sentenças nos documentos, ao invés disso, os autores adotaram a estratégia de ir adicionado iterativamente no resumo a sentença que minimiza o valor da medida KL, até que o tamanho máximo do resumo seja alcançado.

Abordagens baseadas em grafo, como o *TextRank* e *LexRank*, também são muito populares para sumarização multidocumento. Baralis et al. (2013) apresentaram uma abordagem baseada em grafos para sumarização multidocumento que explora regras de as-

sociação para descobrir correlações entre vários termos. A abordagem proposta tem a vantagem de não depender da existência de recursos semânticos, como taxonomias ou ontologias. Primeiro, os documentos são pré-processados e organizados de forma transicional, permitindo assim que regras de associação possam ser descobertas. Então, regras muito frequentes, e que possuem uma alta correlação com termos da coleção de documentos, são extraídas, e um grafo de correlação é gerado a partir desses termos. A medida de Lift (TAN; KUMAR; SRIVASTAVA, 2002) é usada para avaliar o peso das correlações positivas e negativas entre os termos que ocorrem com frequência, e isso reflete a força da associação entre um par de termos. A relevância dos vértices do grafo de correlação é estimada por uma variante do algoritmo *PageRank* (BRIN; PAGE, 1998). Os vértices com maior correlação positiva com outros vértices são colocados no início, enquanto que os vértices com maior correlação negativa com outros vértices são penalizados. O processo de seleção das sentenças é realizado de forma gulosa, selecionando as sentenças que possuem maior cobertura de vértices no grafo de correlação e que também possuem maior pontuação de relevância.

Fattah (2014) propôs uma abordagem supervisionada para a sumarização multidocumento baseada na classificação e seleção de sentenças utilizando técnicas de pontuação de sentenças superficiais. A abordagem proposta utiliza algoritmos de aprendizagem de máquina utilizando as seguintes características para compor o vetor usado para representar cada sentença: **(i)** similaridade entre as sentenças, **(ii)** similaridade entre os parágrafos, **(iii)** explorando a formatação do texto, **(iv)** presença de expressões chave, **(v)** TF-IDF, **(vi)** similaridade das sentenças com o título do documento, **(vii)** posição das sentenças, e **(viii)** presença de expressões não essenciais. Essas características são usadas para treinar três classificadores: *Naive Bayes*, *Maximum Entropy model* e *Support Vector Machine* (SVM). Na abordagem de máxima entropia, uma distribuição de probabilidade uniforme é formada e atribuída em relação às restrições das características. O algoritmo *Naive Bayes* é usado para classificar cada sentença, como importante ou não importante, além disso, cada frase recebe uma pontuação. O algoritmo SVM é usado para obter o hiperplano ideal que separa os exemplos das duas classes. Em seguida, um modelo híbrido de classificação é gerado através da combinação dos três modelos acima. Esse modelo híbrido é aplicado para ranquear as sentenças mais importantes para compor o resumo.

Recentemente, diversas abordagens para a sumarização multidocumento têm adotado a estratégia de modelar o processo de sumarização como um problema de máxima cobertura com restrições, utilizando PLI para resolução desse problema de otimização. Gillick et al. (2009) apresentaram o ICSISumm, um sistema que adota uma abordagem baseada em conceitos utilizando PLI para encontrar o resumo ótimo, ao invés de selecionar as sentenças mais importantes de forma gulosa. O processo de geração dos resumos é executado otimizando a cobertura de conceitos (bigramas) ponderados pela quantidade de documentos em que ele está presente. Além disso, esse sistema explora técnicas para

a compressão de sentenças através de regras usadas para remover fragmentos irrelevantes de uma sentença a partir da sua árvore sintática. Boudin, Mougard e Favre (2015) propuseram o Sume, outro sistema baseado na cobertura de conceitos relevantes para a sumarização multidocumento que estende o modelo de PLI proposto por Gillick et al. (2009) para obtenção de uma única solução, explorando a utilização de limiares para filtrar conceitos com baixo escore de relevância. Assim como o ICSISumm, o sistema Sume utiliza bigramas como conceitos e pondera os pesos desses conceitos utilizando a quantidade de documentos em que o conceito está presente. Cao et al. (2015) propuseram uma abordagem para a sumarização multidocumento utilizando redes neurais recursivas para ranquear sentenças com base na sua importância. O ranqueamento das sentenças é realizado através de um processo de regressão hierárquico que avalia a relevância de uma frase (vértice não-terminal) na árvore de análise. Com base nas supervisões, partindo do nível de palavra para o nível de sentenças, as redes neurais recursivas são usadas para aprender características de classificação sobre a árvore a partir de um vetor de entrada com diversas características. As pontuações de ranqueamento são atribuídas as palavras, que posteriormente são usadas para selecionar frases importantes e não redundantes para formar os resumos. O processo de seleção das sentenças é modelado como um problema de otimização utilizando PLI.

Banerjee, Mitra e Sugiyama (2015b) propuseram uma abordagem abstrativa para a sumarização multidocumento que maximiza a informatividade e a qualidade linguística dos resumos gerados, utilizando um modelo baseado em PLI. A abordagem proposta inicialmente seleciona o documento com maior similaridade do cosseno em relação aos demais documentos do conjunto. Os autores partem da suposição de que o documento selecionado possui as informações mais relevantes que são compartilhadas entre todos os documentos do conjunto. Após esse processo, as sentenças do documento central são usadas para gerar **n** grupos de frases, de modo que cada sentença do documento central inicializa um novo grupo. Para expandir os grupos gerados, é realizada uma verificação da pertinência com cada sentença dos outros documentos do conjunto, excluindo o documento central, com cada grupo de sentenças gerado anteriormente utilizando a similaridade do cosseno. As sentenças são atribuídas ao grupo com o qual elas possuem maior semelhança e caso essa similaridade seja maior que um dado limiar. Para cada grupo de sentenças gerado, um grafo de palavras (FILIPPOVA, 2010) é criado com o objetivo de gerar **M** novas sentenças realizando o processo de fusão das sentenças. Por fim, o processo de seleção de sentenças é modelado como um problema de otimização adotando PLI, com o objetivo de maximizar a informatividade e a qualidade linguística do resumo gerado. Para evitar redundância, apenas uma sentença por grupo pode ser selecionada. O tradicional algoritmo *TextRank* foi adotado para estimar a informatividade das sentenças, e para estimar a qualidade linguística do resumo gerado, um modelo de língua usando trigramas foi adotado para atribuir uma pontuação de qualidade para cada sentença gerada pelo

processo de fusão.

Hong, Marcus e Nenkova (2015) apresentaram uma abordagem para sumarização multidocumento baseada na geração e combinação de diversos resumos candidatos utilizando vários sistemas de sumarização. Nessa abordagem, inicialmente quatro sistemas não supervisionados são utilizados para gerar resumos básicos. Em seguida, esses resumos são combinados no nível de sentenças para gerar novos resumos candidatos. Por fim, os autores utilizam o algoritmo de máquina de vetores de suporte para selecionar o resumo mais representativo entre os resumos candidatos, utilizando uma extensa coleção de características que são usadas para capturar as informações mais importantes a partir de diferentes perspectivas.

Wan et al. (2015) apresentaram uma abordagem discriminativa para a sumarização multidocumento baseada na estratégia de primeiro, gerar diversos resumos e, posteriormente, discriminar o resumo mais informativo. O processo de geração dos resumos candidatos é realizado utilizando um método de sumarização baseado na maximização da cobertura de unigramas no resumo gerado usando PLI. Posteriormente, o algoritmo RankSVM (JOACHIMS, 2002) é usado para ranquear os resumos candidatos, visando selecionar os resumos mais informativos nas primeiras posições do ranque gerado. O algoritmo RankSVM foi treinado usando diversas características extraídas do próprio resumo e também dos documentos de entrada.

Ren et al. (2016) propuseram um arcabouço para a sumarização multidocumento que aplica um modelo de regressão para estimar a importância relativa a uma sentença $f(s|S)$, considerando também o conjunto de sentenças S já selecionadas para o resumo. Para isso, os autores adotam no modelo de regressão, atributos somente da nova sentença s, como posição, tamanho, e também características que envolvem a relação entre essa sentença e outras sentenças S já presentes no resumo. O modelo de regressão é treinado, visando estimar o ganho relativo com a inclusão de uma nova sentença s ao conjunto de sentenças S do resumo em geração, em relação à medida do ROUGE-2. Dessa forma, o processo de seleção das sentenças é realizado de forma gulosa, inserindo a sentença que maximiza a função $f(s|S)$ até que o tamanho máximo do resumo seja alcançado.

Peyrard e Eckle-Kohler (2016) definiram uma abordagem para sumarização multidocumento que maximiza uma aproximação da medida do ROUGE via PLI. Para isso, os autores propõem reduzir o problema de computar a medida do ROUGE, em nível de sentença, usando um algoritmo de regressão, ao invés do resumo como um todo. Os autores exploram técnicas clássicas, como posição, tamanho e similaridade com o título, para representar as frases. Além dessas técnicas, os seguintes métodos de frequência são computados: TF-IDF, a soma das frequências dos bigramas da sentença, a soma dos valores do método de frequência dos documentos (número de documentos que mencionam o n-grama) de todos os unigramas e bigramas da frase.

Ren et al. (2017) apresentaram uma abordagem para a sumarização multidocumento

que considera as relações contextuais entre as sentenças para melhor estimar sua relevância aplicando algoritmos de regressão. Os autores utilizam as relações entre as sentenças com uma rede neural convolutiva, em nível de palavras, para construir as representações das sentenças. Posteriormente, as relações contextuais são adotadas com uma rede neural recorrente, em nível de sentenças, para construir as representações de contexto. Usando esses dois níveis de representação, a solução proposta é capaz de mensurar a relevância das sentenças e suas palavras, no contexto específico de uma determinada frase. O modelo proposto aprende automaticamente, características contextuais relevantes, aprendendo de forma conjunta representações para as sentenças e escores de similaridade entre uma frase e outras sentenças que estão no seu contexto. Finalmente, o processo de seleção das sentenças é realizado de forma gulosa, selecionando a sentença com maior escore gerado pelos algoritmos de redes neurais e que não possua uma alta sobreposição de bigramas com as sentenças já inseridas no resumo.

No Quadro 3 são apresentados de forma resumida os principais trabalhos relacionados, sendo eles caracterizados utilizando as seguintes dimensões: **(i)** Tipo da abordagem de sumarização (Abstrativa e Extrativa); **(ii)** O processo de seleção das sentenças para compor o resumo é iterativo, gulosa ou usa PLI; **(iii)** A abordagem desenvolvida é ou não supervisionada; e **(iv)** Principal vantagem e limitação do trabalho desenvolvido.

2.4 Métodos de Avaliação

O processo de avaliação dos resumos gerados por sistemas de SAT é uma tarefa de extrema importância para o progresso da área (GAMBHIR; GUPTA, 2016; LLORET; PLAZA; AKER, 2017). Avaliar os resumos gerados é essencial para possibilitar a comparação e a replicação dos resultados e, portanto, estimular cada vez mais o desenvolvimento de novas abordagens. Avaliar centenas de resumos manualmente é uma tarefa que demanda tempo e esforço. Dessa forma, é imprescindível o desenvolvimento de medidas e ferramentas de avaliação capazes de realizar esse processo de forma automática ou semiautomática de forma rápida e confiável. Um bom resumo, além de apresentar as informações mais relevantes dos documentos originais, também deve ser coeso, coerente e gramaticalmente correto. Avaliar esses aspectos é uma tarefa complexa e passível de subjetividade. Mesmo para um humano, é difícil identificar quais informações são mais relevantes e merecem estar no resumo gerado.

A avaliação dos resumos pode ser dividida em duas grandes categorias: *Extrínseca* e *Intrínseca* (GAMBHIR; GUPTA, 2016).

Avaliação Extrínseca: Esse tipo de avaliação mensura a qualidade de um resumo com base no seu impacto em outras tarefas, como classificação de texto e recuperação de informação. Dessa forma, um resumo é considerado bom se ele for utilizado com sucesso em outros contextos.

Quadro 3 – Resumo dos principais trabalhos relacionados para a tarefa sumarização multidocumento.

Trabalho	**Abordagem**	**Método de Sumarização**	**Supervisionado**	**Vantagem**	**Limitação**
Nenkova, Vanderwende e McKeown (2006)	Extrativa	Iterativa (sentenças)	-	Requer pouco processamento computacional.	Adota somente a frequência das palavras como medida de relevância.
Conroy, Schlesinger e O'Leary (2006)	Extrativa	Iterativa (sentenças)	-	Requer pouco processamento computacional.	Necessita de um corpus externo para o calculo do teste LLR.
Gillick et al. (2009)	Abstrativa	PLI (bigramas)	-	Adota técnicas de compressão de sentenças.	Usa somente um método para ponderar o peso dos conceitos.
Haghighi e Vanderwende (2009)	Extrativa	Gulosa (sentenças)	-	Explora o método KL para minimizar a divergência entre o resumo em geração e os documentos de entrada.	Explora somente a probabilidade de ocorrência dos n-gramas em conjunto com o método KL.
Baralis et al. (2013)	Extrativa	Gulosa (sentenças)	-	Explora a relação entre as palavras usando uma representação baseada em grafos.	Não considera aspectos importantes como posição das sentenças. Avaliou usando somente um corpus.
Fattah (2014)	Extrativa	Iterativa (sentenças)	Classificação (Sentenças)	Adota uma abordagem hibrida que combina três modelos de aprendizagem de máquina.	Análise somente a relevância das sentenças individualmente.
Banerjee, Mitra e Sugiyama (2015b)	Abstrativa	PLI (sentenças)	-	Adota técnicas de fusão de sentenças.	Negligência outros aspectos relevantes para sumarização, como posição, durante a mensuração da relevância das sentenças. Avaliou usando somente um corpus.

Boudin, Mougard e Favre (2015)	Extrativa	PLI (bigramas)	-	Requer um baixo custo computacional.	Usa somente um método para ponderar o peso dos conceitos.
Cao et al. (2015)	Extrativa	PLI (unigramas, sentenças)	Redes Neurais (sentenças, unigramas)	Explora redes neurais recursivas para ponderar conjuntamente sentenças e unigramas.	Complexidade relativa às redes neurais recursivas. Não considera a importância do resumo como um todo.
Hong, Marcus e Nenkova (2015)	Extrativa	Combinação de resumos (sentenças)	Regressão (resumos)	Estima o resumo mais informativo a partir de diversos resumos candidatos.	Alto custo computacional para gerar os resumos candidatos. Não considera aspectos importantes, como posição e centralidade durante a estimação da informatividade dos resumos candidatos.
Wan et al. (2015)	Extrativa	PLI (unigramas)	Regressão (unigramas) e Ranqueamento (resumos)	Discrimina o resumo mais informativo a partir de diversos resumos candidatos.	Não considera aspectos importantes, como posição e centralidade durante a análise dos resumos candidatos. Avaliou usando somente um corpus.
Ren et al. (2016)	Extrativa	Gulosa(sentenças)	Redes Neurais (sentenças, grupos de sentenças)	Considera a relevância das sentenças individualmente e em conjunto com outras sentenças.	Custo computacional do algoritmo guloso.
Peyrard e Eckle-Kohler (2016)	Extrativa	PLI (sentenças)	Regressão (sentenças)	Visa maximizar uma aproximação da medida do ROUGE das sentenças usadas para compor o resumo.	Somente atributos em nível de n-gramas e sentenças são exploradas. O modelo de regressão é treinado usando um aproximação da medida do ROUGE.
Ren et al. (2017)	Extrativa	Gulosa(sentenças)	Redes Neurais (sentença, palavras)	Explora o contexto no qual a sentença está inserida.	Possui uma alta complexidade relativa aos algoritmos de redes neurais.

Avaliação Intrínseca: Nesse tipo de avaliação, a qualidade e a informatividade dos resumos são avaliadas. A informatividade de um resumo é dada mensurando a intersecção de informações que ele possui com um ou mais resumos de referência. Já a qualidade é avaliada observando aspectos como coesão, coerência e gramaticalidade. Em geral, a informatividade dos resumos é avaliada de forma automática, enquanto que a qualidade é analisada manualmente por avaliadores humanos, em geral através da aplicação de questionários.

Dada a importância do processo de avaliação para o progresso da área de SAT, várias conferências e competições, como a SUMMAC (1996–1998) (MANI et al., 2002), a DUC (2001-2007) (OVER; DANG; HARMAN, 2007) e, mais recentemente, a TAC (OWCZARZAK; DANG, 2011) têm abordado com especial interesse aspectos relacionados à avaliação de sistemas de SAT. As avaliações dos sistemas participantes nas competições dessas conferências foram realizadas manualmente e automaticamente. Essas conferências desempenharam um papel fundamental no desenvolvimento das principais medidas e ferramentas de avaliação usadas atualmente para avaliar sistemas de SAT (GAMBHIR; GUPTA, 2016). Além disso, elas também contribuíram para a realização de meta-avaliações das medidas de avaliação, já que é possível verificar quais medidas de avaliação automática correlacionam melhor com as avaliações humanas.

Em termos gerais, existem três aspectos principais que precisam ser considerados nas avaliações automáticas de resumos (GAMBHIR; GUPTA, 2016): **(i)** determinar quais são as informações mais importantes que devem ser mantidas a partir do texto inicial; **(ii)** reconhecer automaticamente fragmentos de informação no resumo, uma vez que essa informação pode ser expressa de várias formas, por exemplo, utilizando sinônimos; e **(iii)** avaliar a gramaticalidade, coesão e coerência dos resumos.

Os métodos de avaliação de resumos podem ser executados de forma automática, manual ou semiautomática (GAMBHIR; GUPTA, 2016). As abordagens manuais são realizadas por avaliadores humanos que analisam os resumos gerados, observando a informatividade e a qualidade dos resumos. Esse tipo de avaliação parece mais confiável, mas é passível de viés, devido à sua subjetividade. Abordagens automáticas comparam fragmentos de textos a partir do resumo gerado com um ou mais resumos de referência. Essa abordagem é mais rápida de ser executada, mas apresenta problemas relacionados a como verificar a intersecção semântica de informações entre os resumos. Essa limitação é realçada, principalmente quando se comparam resumos extrativos com abstrativos. As abordagens semiautomáticas permitem que um anotador humano manualmente analise, selecione e pondere os fragmentos de informações mais importantes do texto original, e então os resumos gerados automaticamente são ranqueados com base no maior número de fragmentos de informações relevantes que eles possuem.

Diversas abordagens para avaliação de sistemas de SAT têm sido propostas nos últimos anos (LLORET; PLAZA; AKER, 2017). Dentre elas, destacam-se o *Recall-Oriented Unders-*

tudy for Gisting Evaluation (ROUGE) (LIN, 2004) (método automático), o PYRAMID (NENKOVA; PASSONNEAU; MCKEOWN, 2007) (método semiautomático), e vários indicadores que são geralmente usados em métodos de avaliação manual para analisar a qualidade dos resumos. Além disso, existem trabalhos que investigam como avaliar resumos gerados automaticamente quando resumos de referência não estão disponíveis. Uma visão mais abrangente do processo de avaliação de resumos e seus desafios podem ser encontrados em (LLORET; PLAZA; AKER, 2017). No Quadro 4 são apresentadas de forma resumida as abordagens de avaliação supracitadas, e mais detalhes sobre elas são apresentadas nas subseções a seguir.

Quadro 4 – Resumo das abordagens de avaliação de sistemas de SAT.

Abordagem de Avaliação	**Vantagem**	**Limitação**
Manual	Mais precisa e confiável.	Alto custo para realização.
ROUGE	Completamente automática.	Somente realiza uma análise léxica dos resumos.
PYRAMID	Possibilita uma análise mais semântica dos resumos.	Alto esforço manual.
Sem Resumos de Referência	Completamente automática e dispensa a necessidade de resumos de referência.	A análise realizada ainda é muito imprecisa se comparada com as outras abordagens citadas.

2.4.1 Avaliação Manual

A forma mais natural de avaliação que se pode pensar é utilizando diversos avaliadores humanos para analisar a informatividade e a qualidade dos resumos gerados automaticamente (GAMBHIR; GUPTA, 2016). Contudo, avaliar uma grande quantidade de resumos é inviável neste tipo de abordagem, devido ao esforço humano necessário para executá-la. Além disso, existe ainda aspectos relacionados à subjetividade da avaliação, algo que pode ser contornado com a utilização de vários avaliadores por resumo. Por exemplo, nas conferências do DUC, os juízes tinham que avaliar a informatividade dos resumos gerados, ou seja, o quanto de informações mais relevantes do texto original estavam presentes no resumo gerado (OVER; DANG; HARMAN, 2007). Além disso, eles avaliavam, através de questionários, a qualidade dos resumos. Em conferências mais recentes, como a TAC, os juízes tinham que levar em considerações outros aspectos, por exemplo, na avaliação de sistemas de sumarização baseada em consultas, eles tinham que verificar se a consulta de entrada estava sendo respondida nos resumos gerados.

Alguns dos indicadores que são comumente utilizados para definir resumos com alta qualidade são (SAGGION; POIBEAU, 2013): **(i)** Ser sintaticamente correto; **(ii)** Ser coeso e coerente; **(iii)** Ter uma organização lógica; e **(iv)** Não possuir redundâncias. Avaliar esses indicadores automaticamente é muito complexo, principalmente a organização lógica, a coesão e a coerência dos resumos (SAGGION; POIBEAU, 2013). Por isso, em geral, as

avaliações manuais são feitas com base em questionários que apresentam esses indicadores. Nas conferências do DUC e TAC, por exemplo, os avaliadores humanos eram orientados a atribuir uma pontuação que, em geral, variava de 0 a 10, adotando indicadores como: gramaticalidade, redundância, coesão, estrutura e coerência.

2.4.2 ROUGE

Diante do grande esforço necessário para a realização de avaliações manuais, diversas pesquisas foram realizadas para automatizar a avaliação de sistemas de SAT. Nesse sentido, as medidas disponibilizadas pela ferramenta *Recall-Oriented Understudy for Gisty Evaluation* (ROUGE) proposto por Lin (2004) são umas das mais utilizadas para avaliação sistemas de SAT. Essas medidas são baseadas na comparação de uma sequência de elementos (n-grama) entre os resumos gerados automaticamente e um conjunto com um ou mais resumos de referência, geralmente criados manualmente por avaliadores humanos. Em geral, diversos resumos de referência são usados na avaliação, o que possibilita uma maior flexibilidade e confiabilidade na avaliação.

Existem diversas medidas disponíveis na ferramenta ROUGE, sendo as principais delas:

ROUGE-N: Essa medida é baseada na sobreposição de n-gramas, em geral unigramas (N=1) ou bigramas (N=2) são usados, mas trigramas (N=3) e quadrigramas (N=4) também podem ser adotados. Uma sequência de n-gramas é extraída dos resumos candidatos e do conjunto de resumos de referência. A pontuação de cada resumo candidato é dada pela proporção entre a intersecção de n-gramas (do resumo candidato e do conjunto de resumos de referência) e o total de n-gramas extraídos do(s) resumo(s) de referência. A Equação 2.1 apresenta como essa medida é calculada.

$$ROUGE - N = \frac{\sum_{S \in S_{ref}} \sum_{n-grama \in S} Count_{match}(n - grama)}{\sum_{S \in S_{ref}} \sum_{n-grama \in S} Count(n - grama)} \tag{2.1}$$

no qual,

- S_{ref} é o conjunto de resumos de referência;
- n é o tamanho do n-grama;
- $Count_{match}(n - grama)$ é o número máximo de n-gramas co-ocorrendo entre o resumo candidato e o conjunto de resumos de referência;
- $Count(n - grama)$ computa o total de n-gramas que ocorrem nos resumos de referência.

ROUGE-L: Essa medida leva em consideração a maior sequência de palavras em comum entre o resumo gerado e o conjunto de resumos de referência, com lacunas entre as

palavras sendo ignoradas. Por exemplo, dada as frases *The policeman killed the gunman.* e *The policeman chased, shot and killed the terrorist.*, a maior sequência de caracteres entre elas é a cadeia *The policeman killed the.*

ROUGE-W: Essa medida é uma versão ponderada do ROUGE-L, que atribui maior peso às sequências de palavras em comum que possuem mais palavras sucessivas. Por exemplo, dada as frases *S1: The policeman killed the gunman.*, *S2: The policeman chased, shot and killed the terrorist.* e *S3: The policeman killed the terrorist.*, as sentenças *S2* e *S3* possuem o mesmo ROUGE-L (*The policeman killed the*), enquanto que na medida ROUGE-W a sentença *S3* é a melhor opção porque possui mais palavras consecutivas.

ROUGE-S: Essa medida computa a proporção de bigramas não contínuos (bigramas com saltos), em comum entre o resumo gerado e uma coleção de resumos de referência. Por exemplo, a frase *The policeman killed the gunman.* possui 10 bigramas com salto [1], enquanto que a sentença *The policeman chased, shot and killed the terrorist.* possui 28 bigramas com salto. Após a extração, computa-se a interseção entre os bigramas com salto do resumo gerado e dos resumos de referência.

ROUGE-SU4: Essa medida restringe a medida ROUGE-S para considerar somente bigramas com saltos de até no máximo quatro palavras de distância entre os termos do bigrama. Por exemplo, dada a frase *The policeman chased, shot and killed the terrorist.*, o bigrama com salto *policeman terrorist*, considerado na medida do ROUGE-S, não seria considerado na medida do ROUGE-SU4 porque existe mais de quatro palavras entre *policeman* e *terrorist*.

As medidas de avaliação do ROUGE têm sido uma das principais formas de avaliação adotadas na literatura nos últimos anos (SAGGION; POIBEAU, 2013; GAMBHIR; GUPTA, 2016). Contudo, uma importante limitação é o fato de que apenas sobreposições léxicas são consideradas. Diante disso, diversas outras medidas de avaliação têm sido propostas, com o objetivo de proporcionar uma análise mais semântica para computar a similaridade entre os resumos candidatos e os resumos de referência.

2.4.3 PYRAMID

Como mencionado anteriormente, as medidas do ROUGE são baseadas na sobreposição léxica de sequências de n-gramas entre os resumos candidatos e um conjunto de resumos de referência. Uma importante limitação dessas medidas é incapacidade de considerar aspectos semânticos, como a presença de sinônimos, ou caso ocorra uma reformulação das sentenças do texto original. O método PYRAMID proposto por Nenkova,

[1] Esse valor é calculado usando a formula geral de combinações $C(p,2) = \frac{p!}{2!(p-2)!}$, no qual p é total de palavras da sentença.

Passonneau e McKeown (2007) tem por objetivo superar essas limitações, proporcionando uma avaliação mais semântica.

O método PYRAMID é baseado na extração de unidades de conteúdo de sumarização, do inglês *Summarization Content Units* (SCU). As SCU podem ter tamanhos diferentes, variando desde uma única palavra até fragmentos maiores formados por fragmentos de uma sentença. Dessa forma, a primeira etapa consiste na extração manual das SCUs por um grupo de juízes humanos. Então, as SCUs podem ser ponderadas manualmente pelos juízes ou, por exemplo, adotando uma estratégia como atribuir menor peso para SCUs que apareçam em apenas um resumo de referência e maior peso para aquelas que estejam presentes em todos os resumos de referência. A analogia com a palavra Pirâmide é decorrente do fato de que, após o processo de extração, existirão diversas SCUs com uma baixa pontuação na base da pirâmide, e algumas SCUs com uma alta pontuação no topo. Além disso, para cada SCU é associada uma lista de expressões equivalentes (sinônimos). Dessa forma, é possível mapear diversas sequências do documento para uma única SCU. Por fim, é realizada a computação da pontuação dos resumos candidatos com base na quantidade e no peso das SCUs que eles possuem.

Com base na pirâmide de SCUs montada, a informatividade de um novo resumo pode ser definida como a proporção da soma dos pesos das SCUs que ele possui em relação ao total de pesos de um resumo ótimo com o mesmo número de SCUs. Por exemplo, dado um resumo contendo m SCUs, o resumo ótimo considerado é formado pelas mesmas m SUCs presentes nos níveis mais altos da pirâmide.

Esse tipo de avaliação demonstrou ser mais preciso do que as avaliações conduzidas utilizando as medidas do ROUGE (NENKOVA; PASSONNEAU; MCKEOWN, 2007). Nas avaliações conduzidas na conferência do TAC 2008, o método PYRAMID apresentou maior correlação do que as medidas do ROUGE com as avaliações feitas pelos juízes humanos (SAGGION; POIBEAU, 2013). Contudo, esse método requer um grande esforço manual para identificar as SCUs e associar diversas expressões equivalentes para cada uma delas.

2.4.4 Avaliação sem Resumos de Referência

Os métodos de avaliação do ROUGE e do PYRAMID necessitam que resumos de referência estejam disponíveis para que eles possam ser computados. Dado o custo necessário para a criação de resumos de referências, nem sempre eles estão disponíveis. Além disso, em métodos semiautomáticos, como o PYRAMID, ainda existe o custo associado com a extração e as análises das SCUs. Por outro lado, é obvio que o texto de entrada a ser sumarizado sempre está disponível e ele possui informações valiosas que podem ser usadas para avaliar os resumos candidatos gerados.

A ideia de avaliar sistemas de SAT sem a presença de resumos de referência, utilizando próprio conteúdo do documento a ser sumarizado, foi investigada por Louis e Nenkova (2009). Os autores propuseram utilizar as seguintes características extraídas dos resumos

candidatos e do documento original: **(i)** Tamanho do resumo gerado e do texto de entrada; **(ii)** Presença de palavras chave; **(iii)** Propriedades relacionadas à teoria da informação, extraídas a partir da distribuição das palavras do documento; e **(iv)** Similaridade entre os resumos e os documentos. Os autores compararam os resumos candidatos utilizando as quatro características supracitadas mais a medida de divergência *Jensen-Shannon* (LIN, 2006), usada para verificar a divergência entre duas distribuições de probabilidade. Diversos experimentos foram conduzidos utilizando diferentes corpora nas tarefas de sumarização baseada em consultas e sumarização de atualizações. Os resultados experimentais obtidos demonstraram que as características investigadas apresentaram uma alta correlação com o método de avaliação PYRAMID e com o ROUGE.

Outro trabalho nesta mesma linha de investigação foi conduzido por Saggion et al. (2010). Nesse trabalho, os autores demonstraram que a ideia de avaliar sistemas de SAT sem resumos de referência foi eficiente na maioria dos casos, mas em algumas tarefas, como na sumarização de opiniões e de biografias, a performance do método foi baixa. Para suprir esse problema, os autores propuseram um novo método baseado em n-gramas, bigramas com saltos e na medida de divergência *Jensen-Shannon*. Além disso, os autores realizaram diversos experimentos e obtiveram resultados promissores para diferentes tarefas de sumarização em diversos idiomas.

2.5 Considerações Finais do Capítulo

Este capítulo apresentou uma breve introdução aos principais conceitos relacionados a área de SAT, de forma a permitir uma melhor compreensão do restante deste trabalho de doutorado. Para uma visão mais abrangente da área, os seguintes *surveys* são sugeridos: (NENKOVA; MCKEOWN, 2012; LLORET; PALOMAR, 2012; SAGGION; POIBEAU, 2013; GAMBHIR; GUPTA, 2016).

A análise da literatura na tarefa de sumarização monodocumento evidenciou que existem alguns trabalhos (HIRAO et al., 2013; KIKUCHI et al., 2014; PARVEEN; RAMSL; STRUBE, 2015a, 2015a; DURRETT; BERG-KIRKPATRICK; KLEIN, 2016) que modelam o processo de sumarização, como um problema de máxima cobertura, adotando PLI. Contudo, esses trabalhos adotam unidades de discurso ou sentenças como fragmentos textuais que são extraídos e ponderados. No melhor do conhecimento do autor deste trabalho, nenhum dos trabalhos identificado adota uma modelagem baseada em conceitos (unigramas, bigramas) conforme proposto por Gillick et al. (2009). Além disso, a maioria dos trabalhos identificados são estáticos, ou seja, adotam uma abordagem com um conjunto de parâmetros pré-definidos para todos os documentos de entrada. Conforme apontado por Hong, Marcus e Nenkova (2015), na tarefa de sumarização multidocumento, tal característica é uma significante limitação das atuais abordagens extrativas de sumarização.

A revisão da literatura, na tarefa de sumarização multidocumento, ressaltou que essa

área tem sido mais focada nos últimos anos do que a sumarização monodocumento (GAMBHIR; GUPTA, 2016). Em especial, as abordagens baseadas em conceitos usando PLI (GILLICK et al., 2009; LI; QIAN; LIU, 2013; BANERJEE; MITRA; SUGIYAMA, 2015b; BOUDIN; MOUGARD; FAVRE, 2015; LI; LIU; ZHAO, 2015) têm se destacado pelos bons resultados obtidos. Essas abordagens visam maximizar a cobertura de conceitos relevantes no resumo gerado, respeitando o tamanho máximo do resumo desejado. Em sua grande maioria, bigramas têm sido adotados como conceitos, e para mensurar sua relevância são aplicados métodos individuais, como a frequência dos documentos em que o conceito é mencionado, ou adotando algoritmos de regressão para estimar a importância dos conceitos. Esses trabalhos possuem uma importante limitação pois adotam uma configuração (forma de representação e método de ponderação) estática para todos os documentos de entrada. Além disso, com exceção dos trabalhos de Hong, Marcus e Nenkova (2015), Wan et al. (2015), nenhum dos outros trabalhos identificados buscam analisar (em nível de resumo) e selecionar o resumo mais informativo.

Somente na tarefa de sumarização multidocumento, é possível observar que poucos trabalhos (HONG; MARCUS; NENKOVA, 2015; WAN et al., 2015) têm investigado a análise, em nível de resumo, visando discriminar o resumo mais informativo a partir de um conjunto de resumos candidatos. Essa lacuna e as limitações observadas nos dois trabalhos existentes, motivaram a proposta da abordagem combinando PLI e regressão proposta neste trabalho de doutorado.

No próximo capítulo serão apresentados os experimentos conduzidos visando investigar diferentes métodos de pontuação de sentenças e estratégias de combinação para as tarefas de SAT monodocumento e multidocumento. Tal investigação será fundamental para identificar quais aspectos influenciam cada uma das tarefas de sumarização.

3 AVALIANDO TÉCNICAS E ESTRATÉGIAS DE COMBINAÇÃO PARA A PONTUAÇÃO DE SENTENÇAS

Abordagens extrativas para a SAT usualmente são executadas em três etapas principais (NENKOVA; MCKEOWN, 2012): **(i)** Criação de uma representação intermediária; **(ii)** Mensuração da importância de elementos textuais como, por exemplo, n-gramas ou sentenças; e **(iii)** Geração do resumo. Dois importantes aspectos que precisam ser tratados nesse tipo de abordagem são: **(i)** Como mensurar a importância de uma sentença; e **(ii)** Como evitar redundância no resumo gerado.

Diversas técnicas para mensurar a relevância das sentenças em abordagens de SAT extrativa têm sido propostas e avaliadas na literatura ao longo dos anos. Essas abordagens podem ser categorizadas quanto a complexidade das técnicas adotadas em: Superficiais e Profundas. As abordagens superficiais são simples de implementar, exigem um baixo processamento computacional e necessitam de pouco ou nenhum conhecimento linguístico. Por outro lado, as abordagens profundas se caracterizam pelo uso de recursos semânticos, como Ontologias (BARALIS et al., 2013), analisadores semânticos (KHAN; SALIM; KUMAR, 2015), analisadores de discurso(JORGE; PARDO, 2010; MAZIERO; JORGE; PARDO, 2014), entre outras técnicas que buscam proporcionar uma maior compreensão do texto.

Nem sempre os resultados alcançados pelas abordagens profundas valem o esforço gasto, dado o custo computacional requerido e a indisponibilidade de recursos semânticos necessários. Por exemplo, em Baralis et al. (2013), os autores fazem uso da ontologia Yago (SUCHANEK; KASNECI; WEIKUM, 2007) para identificar entidades, por exemplo, nome de pessoas, cidades, organizações, mencionadas no texto. O acesso e a disponibilidade de recursos como esse ainda é complexo, e demanda um alto processamento computacional. Diante disso, as técnicas superficiais podem representar uma solução viável para equilibrar os aspectos de desempenho e recursos necessários para execução.

Diversos trabalhos identificaram e avaliaram o desempenho de várias técnicas superficiais para mensurar a importância das sentenças para SAT (NETO; FREITAS; KAESTNER, 2002; BINWAHLAN; SALIM; SUANMALI, 2009; ABUOBIEDA et al., 2013; MEENA; GOPALANI, 2014; FERREIRA et al., 2013; MEENA; DEOLIA; GOPALANI, 2015; SILVA et al., 2015b). Esses trabalhos investigaram a aplicação dessas técnicas individualmente ou utilizando estratégias para combiná-las, por exemplo, usando algoritmos de aprendizagem de máquina (NETO; FREITAS; KAESTNER, 2002; SILVA et al., 2015b), algoritmos evolucionários (BINWAHLAN; SALIM; SUANMALI, 2009; ABUOBIEDA et al., 2013), combinações das pontuações individuais de cada técnica (MEENA; GOPALANI, 2014; FERREIRA et al., 2013; MEENA;

DEOLIA; GOPALANI, 2015), entre outras.

Algumas lacunas podem ser apontadas nos trabalhos citados anteriormente: **(i)** Experimentos conduzidos, na maioria dos casos, utilizando apenas um único corpus por domínio (notícias, blogs, artigos científicos, entre outros), o que compromete a generalização das conclusões obtidas; **(ii)** Alguns trabalhos utilizaram apenas um subconjunto de documentos de um corpus, por exemplo, Meena e Gopalani (2014), Meena, Deolia e Gopalani (2015) utilizaram apenas cem documentos do corpus do DUC 2002 para investigar diversas técnicas superficiais e combinações na tarefa de sumarização monodocumento; **(iii)** Com exceção de Meena, Deolia e Gopalani (2015), que analisaram todas as possíveis combinações de seis métodos de pontuação de sentenças, outros trabalhos não deixaram claro quais os critérios adotados para compor as combinações investigadas; e **(iv)** Falta de uma comparação entre diferentes estratégias para a combinação das técnicas de pontuação.

Diante dessas lacunas, este capítulo tem por objetivo investigar a performance de dezoito técnicas superficiais para mensurar a importância das sentenças nas tarefas de sumarização monodocumento e multidocumento no contexto de artigos de notícias escritos em Inglês. As técnicas investigadas foram selecionadas por serem frequentemente citadas na literatura e por terem apresentado bons resultados em diversos trabalhos. Além disso, quatro estratégias para a combinação das técnicas investigadas foram analisadas.

As principais contribuições deste capítulo são:

- Uma extensa investigação de diversas técnicas superficiais para computar a importância das sentenças e estratégias de combinação considerando as tarefas de sumarização monodocumento e multidocumento. Tal investigação foi conduzida utilizando os corpora do DUC 2001-2002 e o corpus CNN para a sumarização monodocumento, enquanto que para a sumarização multidocumento foram adotados os corpora do DUC 2001-2004.

- Identificação de combinações que apresentam resultados competitivos com diversos sistemas do estado da arte, tanto na sumarização monodocumento quanto na multidocumento.

- As diversas análises realizadas permitiram uma investigação do comportamento das técnicas e sistemas de SAT, possibilitando identificar lacunas que precisam ser resolvidas e quais aspectos influenciam cada tarefa (monodocumento e multidocumento).

O restante deste capítulo está organizado como segue: Na Seção 3.1 são apresentadas uma visão geral do processo de sumarização adotado, e uma breve descrição de cada uma das técnicas de pontuação de sentenças investigadas. Na Seção 3.2 são apresentados os resultados dos experimentos realizados. Por fim, na Seção 3.3 são apresentadas as considerações finais do capítulo.

3.1 Processo de Sumarização Mono e Multidocumento Adotado

O processo de sumarização adotado neste capítulo é composto por três etapas, conforme ilustrado na Figura 2 e brevemente descrito a seguir:

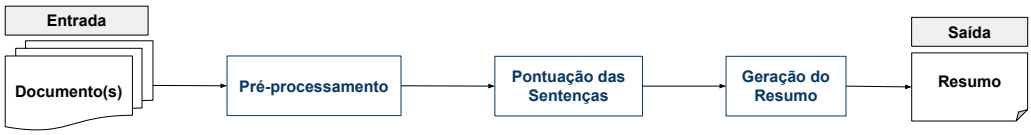


Figura 2 – Visão geral do processo de sumarização adotado.

Pré-processamento: O documento ou o conjunto de documentos de entrada é pré-processado utilizando tradicionais tarefas de PLN. Para tal, a ferramenta *Stanford Natural Languague Processing Toolkit* (CoreNLP) (MANNING et al., 2014) foi utilizada para realizar as tarefas de tokenização, segmentação das sentenças, lematização, etiquetagem das classes gramaticais, reconhecimento de entidade nomeadas e análise sintática.

Pontuação das Sentenças: Nesta etapa, as técnicas de pontuação das sentenças são aplicadas para analisar cada sentença do(s) documento(s) de entrada e gerar um valor que deve refletir qual a sua relevância para ser ou não inserida no resumo. As técnicas analisadas neste capítulo são descritas nas próximas subseções. Todas as técnicas são normalizadas e seus valores variam entre 0 e 1.

Geração do Sumário: As sentenças com maior pontuação são iterativamente inseridas no resumo. Para evitar redundância, utilizou-se a heurística de que uma nova sentença só é inserida no resumo caso ela não possua similaridade do cosseno maior que 0,5 com cada uma das sentenças já incluídas no resumo (NENKOVA; VANDERWENDE; MCKEOWN, 2006). Esse processo é repetido até que não se tenha mais sentenças a inserir ou que a taxa de compressão do resumo seja atingida.

Por fim, uma questão fundamental que precisa ser tratada é a ordenação das sentenças no resumo gerado. Durante a etapa de pré-processamento, caso os documentos de entrada não possuam um código identificador numérico, cada um deles receberá um identificador com base na ordem em que eles são analisados. Além disso, todas as sentenças de um documento são numeradas conforme sua ordem de aparição. Na sumarização monodocumento, o índice da posição de aparição das sentenças no documento de entrada é usado para ordenar as sentenças no resumo. Como cada índice é único em um documento, eles podem ser usados para ordenar as sentenças sem nenhum conflito.

Por outro lado, na sumarização multidocumento, a ordenação das sentenças no resumo gerado não é tão trivial de ser resolvido. Isso acontece porque as sentenças

selecionadas podem pertencer a documentos diferentes, e possuírem o mesmo índice de posição. Para solucionar essa questão, a seguinte estratégia foi adotada: Primeiro, as frases são agrupadas por documento e, em seguida, cada grupo de sentenças é ordenado usando a posição das frases em cada documento. Por fim, os grupos de sentenças são ordenados com base no identificador que o documento recebeu na etapa de pré-processamento. Para ilustrar a estratégia acima imagine que um resumo foi gerado usando um conjunto de sentenças $s_i^{d_j}$, no qual i é o identificador da sentença e j é o do documento. As sentenças selecionadas foram: $s_1^{d_3}$, $s_1^{d_1}$, $s_2^{d_2}$, e $s_4^{d_3}$. Após o processo de ordenação, as sentenças serão apresentadas na seguinte ordem: $s_1^{d_1}$, $s_2^{d_2}$, $s_1^{d_3}$, e $s_4^{d_3}$.

No restante desta seção são apresentadas as dezoito técnicas superficiais para a pontuação de sentenças identificadas na literatura e investigadas neste capítulo.

3.1.1 Técnicas de Pontuação de Sentenças Investigadas

Uma das principais decisões a se tomar nas abordagens extrativas é como mensurar o quão relevante uma sentença é para o entendimento global dos tópicos discutidos no(s) documento(s). Diversas técnicas superficiais para computar a importância de uma sentença para SAT têm sido propostas e avaliadas ao longo dos anos (LUHN, 1958; EDMUNDSON, 1969; NENKOVA; MCKEOWN, 2012; BARRERA; VERMA, 2012; FERREIRA et al., 2013; MEENA; DEOLIA; GOPALANI, 2015). Essas técnicas são baseadas em informações estatísticas, e linguísticas até um certo nível de semântica.

Diferentes níveis de representação (FERREIRA et al., 2013) são adotados como, por exemplo: **(i)** *Nível de Palavras*, no qual cada palavra recebe uma pontuação e as sentenças são pontuadas com base no escore das palavras que elas possuem; **(ii)** *Nível de Sentenças*, utilização de características da própria sentença, como posição, centralidade, semelhança com o título do documento; e **(iii)** *Baseada em Grafo*, explorando a relação entre as sentenças ou entre as palavras usando algoritmos baseados em grafos.

Nas subseções a seguir são apresentadas as técnicas de ponderação de sentenças investigadas neste capítulo, agrupadas nos níveis de representação mencionados anteriormente.

Nível de Palavras

Os primeiros métodos de SAT eram baseados na ideia de que sentenças relevantes são compostas por palavras importantes (LUHN, 1958). Neste tipo de técnica, cada palavra recebe uma pontuação, que deve refletir qual a sua importância para ser incluída no resumo, e a pontuação de uma sentença é dada pela soma da relevância de todas as suas palavras. Os métodos de pontuação investigados são descritos nas subseções a seguir.

Coocorrência das Palavras

O método de coocorrência das palavras (GUPTA; PENDLURI; VATS, 2011; FERREIRA et al., 2013) mensura a coocorrência de um ou mais n-gramas de uma sentença em outras sentenças do(s) documento(s). A ideia deste método é que sentenças que possuem muitas palavras que geralmente ocorrem juntas são mais importantes, já que palavras frequentes podem se referir a relevantes termos do documento. Este método pode ser computado usando a similaridade dos n-gramas de uma sentença em relação as outras sentenças. A Equação 3.1 demonstra como é calculada a pontuação de uma sentença com base neste método.

$$CoocorrenciaPalavras(s_i) = \sum_{j=1, i \neq j}^{S} simNgram(s_i, s_j) \tag{3.1}$$

no qual,

- S é o total de sentenças do(s) documento(s);
- *sim* retorna à similaridade dos n-gramas entre duas sentenças s_i e s_j;
- N é o tamanho dos n-gramas.

Frequência das Palavras

O método de frequência das palavras é uma das técnicas mais antigas aplicadas para mensurar a pontuação de uma sentença para SAT (LUHN, 1958). Este método é baseado na ideia de que sentenças importantes possuem palavras frequentes, ou seja, quanto mais frequente uma palavra é, mais importante ela é para representar as informações principais de um documento. Nem todas as palavras são levadas em consideração, geralmente, *stopwords* são removidas e o algoritmo de *stemming* ou lematização é aplicado antes de computar a frequência das palavras.

Alguns trabalhos (SUANMALI; SALIM; BINWAHLAN, 2009; ABUOBIEDA et al., 2012) demonstraram que levar em consideração apenas as N palavras mais frequentes do(s) documento(s) pode produzir melhores resultados do que considerar todas as palavras. A pontuação de uma sentença com base neste método é calculada conforme apresentado na Equação 3.2.

$$FrequenciaPalavras(s_i) = \sum_{j=1}^{P} freq(p_j) \tag{3.2}$$

no qual,

- *freq* retorna à frequência de uma palavra $p_j \in s_i$ no(s) documento(s) de entrada;
- P é o total de palavras da sentença s_i.

Frequência do Termo - Frequência Inversa das Sentenças

A técnica de Frequência do Termo - Frequência Inversa das Sentenças (FT-FIS) é uma adaptação do tradicional método de *Term Frequency - Inverse Document Frequency* (TF-IDF). A mudança na nomenclatura é dada em alguns trabalhos já que na SAT considera-se o nível de sentenças ao invés de documento. A FT-FIS de uma palavra é computada como apresentado na Equação 3.3 e a pontuação de uma sentença com base neste método é dada como demonstrado na Equação 3.4.

$$FT - FIS(t_i) = FT(t_i) \ \times \ log\left(\frac{S}{Oc_{t_i}}\right) \tag{3.3}$$

$$FT - FIS(s_i) = \sum_{t_j \in T}^{S} FT - FIS(t_j, s_i) \tag{3.4}$$

no qual,

- FT retorna a frequência de um termo t_j no(s) documento(s);
- S é o total de sentenças do(s) documento(s);
- T é o total de termos (palavras) em s_i;
- Oc_{t_i} é o total de sentenças que possuem o termo t_i.

Nível de Sentenças

Esse tipo de técnica é uma das mais antigas na area de SAT, sendo utilizada pela primeira vez por Edmundson (1969). Métodos com granularidade em nível de sentenças analisam características da própria frase, tais como posição, centralidade, presença de certos tipos de elementos, entre outros aspectos. Nas subseções a seguir são apresentas as técnicas investigadas.

Centralidade das Sentenças

A centralidade de uma sentença pode ser definida como o grau de sobreposição entre uma sentença s_i e as outras sentenças do(s) documento(s) (FATTAH; REN, 2009; FERREIRA et al., 2013). Esse método é baseado na hipótese de que sentenças que compartilham muitas informações com outras sentenças descrevem melhor o conteúdo mais importante de um documento. Baseado nisso, a pontuação de uma sentença s_i é computada como demonstrada na Equação 3.5.

$$Centralidade(s_i) = \frac{S_{s_i} \cap S_{s_o}}{S_{s_i} \cup S_{s_o}} \tag{3.5}$$

no qual,

- S_{s_i} é o conjunto de palavras da sentença s_i;
- S_{s_o} é o conjunto de palavras das outras sentenças do(s) documento(s).

Dados Numéricos

Este método é baseado na premissa de que sentenças contendo dados numéricos possuem uma maior probabilidade de serem inseridas nos resumos (FERREIRA et al., 2013). Usualmente, dados números se referem a importantes informações como, datas, porcentagens, valores monetários, entre outros. A pontuação de uma sentença s_i baseado neste método é dado conforme apresentado na Equação 3.6.

$$DadosNumericos(s_i) = \frac{total\_de\_dados\_numericos\_em\_s_i}{P} \tag{3.6}$$

no qual,

- P é o total de palavras de s_i.

Entidades Nomeadas

Reconhecimento de Entidades Nomeadas é a tarefa de identificar e classificar entidades no texto e associá-las com categorias semânticas. Entidades Nomeadas usualmente se referem a nomes de pessoas, lugares, organizações, entre outros. Tais entidades são importantes pois descrevem relevantes informações mencionadas no(s) documento(s). A ideia deste método é que o resumo deve incluir o maior número possível de entidades mencionadas no texto. A pontuação de uma sentença s_i usando esse método é dado como ilustrado na Equação 3.7.

$$EntidadesNomeadas(s_i) = \frac{\#total\_de\_entidades\_em\_s_i}{\#maximo\_de\_entidade\_em\_uma\_sentenca} \tag{3.7}$$

Expressões Chave

A presença de expressões chave (*Cue-phrases*) (EDMUNDSON, 1969) foi um dos primeiros métodos usados para mensurar a importância de uma sentença para SAT. A ideia deste método é que sentenças que possuem muitas expressões que são comumente usadas em resumos gerados por humanos, por exemplo, “Em suma”, “Conclui-se”, “Em síntese”, “Em resumo”, "Para concluir”, “Neste relatório”, entre outras, possuem uma alta probabilidade de serem incluídas no resumo. Esse método necessita da definição prévia de um dicionário contendo diversas expressões que são descobertas a partir da análise de vários

resumos gerados por humanos. Na Equação 3.8 é apresentada como a pontuação de uma sentença é calculada usando essa técnica.

$$ExpressoesChave(s_i) = \frac{total\_de\_expressoes\_chave\_em\_s_i}{total\_de\_expressoes\_chave\_no(s)\_documento(s)} \quad (3.8)$$

Nomes Próprios

Nomes próprios podem se referir a entidades como pessoas, lugares, organizações, entre outros. Sentenças que contém muitos nomes próprios podem ser consideradas importantes, e assim terem uma maior probabilidade de serem inseridas no resumo do documento (FERREIRA et al., 2013). Essa heurística é similar a frequência de entidade nomeadas, a diferença é que ela não é dependente de uma ferramenta que reconheça entidades nomeadas. A pontuação de uma sentença com base nessa heurística é calculada conforme apresentada na Equação 3.9.

$$NomesProprios(s_i) = \frac{total\_de\_nomes\_proprios\_em\_s_i}{total\_de\_palavras\_em\_s_i} \quad (3.9)$$

Palavras com Inicias Maiúsculas

Palavras que iniciam com letras maiúsculas, geralmente, podem ser consideradas mais relevantes, pois, usualmente podem se referir a importantes termos como acrônimos, nomes de pessoas, lugares, organizações, entre outros (FERREIRA et al., 2013). Essa heurística atribui uma maior pontuação para sentenças que contém mais palavras com inicias maiúsculas. A pontuação de uma sentença com base nessa heurística é calculada como apresentado na Equação 3.10.

$$PalavrasMaiusculas(s_i) = \frac{P_m}{P} \quad (3.10)$$

no qual,

- P_m é o total de palavras com letras maiúsculas na sentença s_i. Palavras como artigos, determinantes e preposição não são levadas em consideração;
- P é o total de palavras da sentença s_i.

Posição das Sentenças

A posição de uma sentença no documento é uma das heurísticas que apresenta melhor performance para a SAT, principalmente em artigos de notícias (EDMUNDSON, 1969; OUYANG et al., 2010; FERREIRA et al., 2013). Essa heurística é baseada na ideia de que as primeiras sentenças de um documento são as mais relevantes, e que sua importância para

o resumo diminui à medida que ela se afasta do início. Alguns trabalhos propõem atribuir maior pontuação a sentenças no início e também no fim do documento (FERREIRA et al., 2013). Para documentos longos, como artigos científicos ou livros, a posição da sentença pode ser contada a cada novo parágrafo.

Ferreira et al. (2013) atribui uma maior pontuação para sentenças no início e no fim do documento utilizando a seguinte estratégia: a primeira sentença recebe a pontuação $\frac{N}{N}$, a segunda sentença recebe $\frac{N-1}{N}$, e assim por diante, no qual **N** é um limiar para o número de sentenças que devem ser levadas em consideração. A mesma ideia é aplicada, mas agora iniciando no fim do documento.

Abuobieda et al. (2012), por outro lado, usa a estratégia de atribuir uma maior pontuação para sentenças apenas no início do documento. O escore de uma sentença baseada nessa estratégia é computada conforme apresentado na Equação 3.11.

$$PosicaoSentenca(s_i) = 1 - \frac{i}{S} \tag{3.11}$$

no qual,

- i é índice da posição da sentença no documento, com i começando em 0;
- S é o total de sentenças do documento.

Relações Abertas

Extração de Relações Abertas, do inglês *Open Information Extraction* (OIE) (ETZIONI et al., 2011; MAUSAM et al., 2012), consiste no processo não supervisionado de extração de relações binárias no formato relação(argumento1, argumento2) a partir de um texto de entrada. Por exemplo, dada a sentença "*Maurice Levy is the head of the one of the world's largest communication firms*", um sistema de OIE extrai a relação "*is the head of(Maurice Levy, one of the world's largest communication firms)*". A presença de várias relações em uma sentença pode ser um importante indício de que essa frase descreve fatos importantes que merecem ser inseridos no resumo. Baseado nessa ideia, esse método atribui maior importância a sentenças que possuem mais relações. A Equação 3.12 ilustra como é calculada a pontuação de uma sentença s_i usando esse método.

$$RelacoesAbertas(s_i) = \frac{RA_{s_i}}{Max_{RA}} \tag{3.12}$$

na qual,

- RA_{s_i} é o total de relações abertas extraídas na sentença s_i;
- Max_{RA} é o total de relações abertas extraídas no(s) documento(s).

Similaridade com o Título

O título de um documento geralmente apresenta indícios do principal tema abordado no documento, principalmente em artigos de notícias. A ideia dessa heurística é que sentenças que possuem uma maior similaridade com o título do documento são as mais importantes e por isso devem ser inseridas no resumo (EDMUNDSON, 1969; FERREIRA et al., 2013). Para documentos sem título, alguns trabalhos consideram a primeira sentença do documento como tal. A pontuação de uma sentença com base nessa heurística é calculada como apresentada na Equação 3.13.

$$SimilaridadeTitulo(s_i) = \frac{P_{s_i} \cap P_t}{P_{s_i}} \tag{3.13}$$

na qual,

- P_{s_i} é o conjunto de palavras da sentença s_i;
- P_t é o conjunto de palavras do título do documento.

Similaridade Léxica

Similaridade Léxica (*Lexical Similarity*) (FERREIRA et al., 2013) é baseada na ideia de que as sentenças mais importantes para serem incluídas no resumo são formadas por cadeia de palavras altamente conectadas, ou seja, aquelas que compartilham algum relacionamento semântico como, sinonímia, hiperônimos, hipônimos, ou que estão no mesmo contexto. Baseado nisso, esse método calcula a similaridade entre as palavras de uma sentença, e atribui um peso maior para sentenças que possuem palavras muito conectadas entre si. É apresentada na Equação 3.14 como a pontuação de uma sentença s_i é computada utilizando esse método.

$$SimilaridadeLexica(s_i) = \sum_{j=1,k=1,j\neq k}^{P} sim(p_j, p_k) \tag{3.14}$$

na qual,

- P é o total de palavras da sentença s_i;
- *sim(p_j, p_k)* retorna a similaridade entre as palavras p_j e p_k. Em geral, medidas de similaridade baseadas em algum dicionário léxico como o *Wordnet* (MILLER, 1995) são usadas.

Sintagmas Nominais e Verbais

Sintagmas consistem em uma sequência de elementos léxicos formados por uma unidade central, chamada de núcleo, e seus modificadores. Existem diversos tipos de sintagmas, porém, neste método, apenas os sintagmas Nominais e Verbais serão considerados. Sintagma Nominal, do inglês *Noun Phrase* (NP), é um grupo de substantivos e seus modificadores, cujo o núcleo é um substantivo. Por outro lado, Sintagma Verbal, do inglês *Verbal Phrase* (VP), é formado por um verbo principal e seus complementos. Por exemplo, na frase "*John Snow is going to 2014 FIFA World Cup Brazil.*", os fragmentos "*John Snow*" e "*2014 FIFA World Cup Brazil*" são extraídos como sintagmas nominais e o fragmento "*is going to*" é extraído como um sintagma verbal. Esse método é baseado na ideia de que os sintagmas nominais e verbais podem identificar importantes entidades e ações, respectivamente. A pontuação de uma sentença s_i baseada nesse método é dada conforme apresentada na Equação 3.15.

$$Sintagmas(s_i) = \frac{SN\_SV_{s_i}}{SN\_SV} \tag{3.15}$$

na qual,

- $SN\_SV_{s_i}$ é o total de sintagmas nominais e verbais da sentença s_i;
- SN_SV é o total de sintagmas nominais e verbais do(s) documento(s).

Tamanho das Sentenças

Selecionar sentenças muito pequenas ou muito grandes para compor o resumo pode não ser uma boa estratégia. Sentenças muito pequenas não agregam informações relevantes para representar os principais assuntos discutidos em um documento (FATTAH; REN, 2009; FERREIRA et al., 2013). De maneira similar, selecionar sentenças muito grandes pode ser um desperdício de espaço, já que sentenças geralmente apresentam informações relevantes em parte e detalhes que são irrelevantes para o resumo em outra parte. Diante disso, esse método é executado em duas partes: **(i)** sentenças menores ou maiores que um dado limiar são removidas, e então **(ii)** as sentenças restantes recebem a pontuação de importância com base na Equação 3.16.

$$TamanhoSentenca(s_i) = \frac{\#total\_de\_palavras\_em\_s_i}{\#total\_de\_palavras\_da\_maior\_sentenca} \tag{3.16}$$

Técnicas baseadas em grafos

Nas abordagens baseadas em grafos, a pontuação de uma sentença é dada pelo seu relacionamento com outras sentenças. Por exemplo, quando uma sentença se refere a

outra, pode-se gerar uma aresta entre essas duas sentenças, com um peso associado. Tais pesos podem posteriormente ser utilizados para identificar quais são as sentenças mais relevantes do documento. As técnicas baseadas em grafos analisadas neste capítulo são descritas nas subseções a seguir.

Bushy Path

Bushy Path (SALTON et al., 1997) é baseada na ideia de que sentenças importantes possuem muitas informações compartilhadas com outras sentenças. Esta técnica mensura a importância de uma sentença s_i, que é representada como um vértice em um grafo, computando o total de arestas que a sentença possui no grafo. Uma aresta entre duas sentenças é criada quando a similaridade do cosseno, ou qualquer outra medida de semelhança, entre duas frases é maior que um dado limiar. A ideia é que sentenças altamente conectadas podem representar informações centrais que indicam o principal tópico discutido no(s) documento(s). Na Equação 3.17 é apresentada como a pontuação de uma sentença s_i é calculada usando esta técnica.

$$BushyPath(s_i) = grau(s_i) \tag{3.17}$$

no qual,

- *grau(s_i)* retorna o número de arestas que a sentença s_i possui no grafo.

Similaridade Agregada

A Similaridade Agregada (*Aggregate Similarity*) (JUNG; KO; SEO, 2005; FERREIRA et al., 2013) é um método baseado em grafos que utiliza a ideia de centralidade das sentenças. Esta técnica é muito semelhante a *Bushy Path*, a diferença é que a pontuação de uma sentença s_i é dada pelo somatório dos pesos das arestas de cada vértice (sentença) no grafo. O peso de uma aresta a_{ij} é dado pela similaridade do cosseno entre as sentenças s_i e s_j. Na Equação 3.18 é apresentada como a pontuação de uma sentença s_i é calculada usando esta técnica.

$$SimilaridadeAgregada(s_i) = \sum_{j=1, i \neq j}^{S} peso\_aresta(s_i, s_j) \tag{3.18}$$

no qual,

- S é o total de sentenças no documento ou no conjunto de documentos;
- *peso_arestas* retorna a similaridade entre as sentenças s_i e s_j.

TextRank

TextRank (MIHALCEA; TARAU, 2004) é um tradicional algoritmo baseado em grafos para extração de palavras-chave. Esse método atribui uma alta pontuação para sentenças que possuem muitos n-gramas relevantes. A importância de um n-grama é atribuída pela sua frequência e coocorrência com outros n-gramas frequentes. Barrera e Verma (2012) identificou melhores resultados quando utilizou somente n-gramas formados por substantivos e adjetivos.

O peso de um n-grama é computado conforme apresentado na Equação 3.19, enquanto que a pontuação de uma sentença baseada no TextRank é dada conforme apresentada na Equação 3.20.

$$TextRank(v_i) = (1-d) + d \times \sum_{v_j \in In(v_i)} \frac{w_{ji}}{\sum_{v_k \in Out(v_j)} w_{jk}} \times TextRank(v_j) \tag{3.19}$$

$$TextRank(s_i) = \sum_{j=1, t_j \in s_i}^{P} TextRank(t_j) \tag{3.20}$$

no qual,

- d é o fator de amortecimento que usualmente é definido com o valor de 0,85 (BRIN; PAGE, 1998);
- $In(v_i)$ é o conjunto de vértices que apontam para v_i;
- $Out(v_j)$ é o conjunto de vértices que v_j aponta;
- P é o total de palavras ou n-gramas da sentença s_i;
- t_j é uma palavra ou n-grama pertencente a sentença s_i;
- w_{ji} é o total de coocorrência entre v_i e v_j.

3.2 Experimentos

Nesta seção são apresentados os experimentos conduzidos com o objetivo de avaliar as técnicas de pontuação de sentenças superficiais apresentadas na seção anterior. Para alcançar esse objetivo, várias avaliações foram realizadas: **(i)** Avaliação individual de cada uma das técnica de pontuação de sentenças investigada (Subseção 3.2.3) **(ii)** Análise comparativa entre quatro estratégias para combinar os métodos de pontuação (Subseção 3.2.4); **(iii)** Investigação da utilização de algoritmos de aprendizagem de máquina para a tarefa de classificação e pontuação das sentenças (Subseção 3.2.5); e **(iv)** Comparação entre as melhores configurações encontradas, técnica individual, método de combinação e algoritmo de aprendizagem de máquina, com diversos sistemas do estado da arte (Subseção 3.2.6).

Antes de apresentar os resultados dos experimentos realizados, uma breve introdução do ambiente experimental adotado é apresentada na Subseção 3.2.1, e alguns detalhes das implementações dos métodos de pontuação das sentenças são apresentados na Subseção 3.2.2.

3.2.1 Configurações dos Experimentos

Todos os experimentos foram realizados no contexto das tarefas de sumarização genérica monodocumento e multidocumento no domínio de artigos de notícias escritos em Inglês. Na sumarização monodocumento foram usados os corpora do DUC 2001-2002 e CNN. Na sumarização multidocumento os corpora do DUC 2001-2004 foram adotados. Na Tabela 1 são apresentadas algumas estatísticas básicas desses corpora.

Tabela 1 – Estatísticas dos corpora do CNN e do DUC adotados nos experimentos.

Corpus	#Grupos	#Documentos	#Sentenças	#Palavras	Tarefas
CNN	0	3.000	115.649	2.628.336	Mono
DUC 2001	30	308	11.026	269.990	Mono e Multi
DUC 2002	59	533	14.370	348.012	Mono e Multi
DUC 2003	30	298	7,691	197.483	Multi
DUC 2004	50	500	13.135	336.073	Multi

Os corpora do DUC 2001-2002 possuem 309 e 576 documentos, respectivamente. Contudo, observou-se a ocorrência de documentos duplicados em grupos diferentes. Por isso, na tarefa da sumarização monodocumento, esses documentos repetidos foram removidos, resultando em 308 e 533 documentos distintos nos corpora do DUC 2001 e 2002, respectivamente.

Para avaliar os resumos gerados, foram adotadas duas medidas de avaliação:

- *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) (LIN, 2004): Como sugerido por Hong et al. (2014) a medida de cobertura do Rouge-1 possui uma maior precisão para identificar resumos informativos. Lin (2004) demonstrou que a cobertura do Rouge-2 possui uma alta correlação com avaliações realizadas manualmente por avaliadores humanos. Por isso, nos experimentos conduzidos as medidas de cobertura do ROUGE-1 e ROUGE-2 usando o algoritmo de *stemming* e não removendo as *stopwords*. Essa configuração apresentou a maior concordância com diversas avaliações realizadas por avaliadores humanos em Owczarzak et al. (2012). O ROUGE-1.5.5 foi utilizado com a seguinte linha de comando: *-n 2 -m -f A*, nos quais:

  - **-n 2**: Especifica a quantidade máxima de n-gramas que serão computados. Como n foi definido como dois, serão avaliadas as medidas usando unigramas e bigramas.
  - **-m**: Indica que o algoritmo de *stemming* de Porter (PORTER, 1997) será adotado.
  - **-f A**: Define que caso exista mais de um resumo de referência, a pontuação final será a média aritmética da avaliação individual com cada um desses resumos de referência.
  - Como os modelos de avaliação disponíveis nos corpora do DUC foram construídos usando limiares baseados na contagem de palavras, para estes corpora o parâmetro *-l N* foi usado para truncar o número de palavras levadas em consideração no resumo gerado em N palavras.

- *Intersecção de Sentenças* (IS) (MANI, 2001; FERREIRA et al., 2013): Essa medida computa a intersecção de sentenças entre os resumos gerados automaticamente e o conjunto de resumos de referência. Um importante aspecto dessa medida é a sua habilidade de identificar métodos com uma alta precisão em reconhecer boas sentenças para compor os resumos. Contudo, essa medida só pode ser computada quando resumos extrativos são disponíveis. Por isso, ela foi executada particularmente no corpus CNN. Na Equação 3.21 é demonstrada como a medida IS é computada.

$$IS(r_i) = \frac{S_{resumo}}{S_{referencias}} \tag{3.21}$$

  no qual,

  - S_{resumo} é o total de sentenças do resumo r_i que estão presentes no(s) resumo(s) de referência;
  - $S_{referencias}$ é o total de sentenças presentes no(s) resumo(s) de referência.

Os testes estatísticos realizados nos experimentos seguiram os seguintes passos: **(i)** Primeiro realizou-se o teste de *Shapiro-Wilk* (SHAPIRO; WILK, 1965) para verificar a normalidade da distribuição dos valores da cobertura do ROUGE-1 e ROUGE-2; **(ii)** Se a distribuição segue a normalidade, o teste *T-Student* pareado (GIBBONS; CHAKRABORTI, 2003) é executado, caso contrário, o teste de *Wilcoxon signed-rank* (GIBBONS; CHAKRABORTI, 2003) é executado; e **(iii)** O teste selecionado na etapa anterior é executado duas vezes: primeiro usando a hipótese nula ($M_1 = M_2$) e caso ($p - value < 0.05$) (5% de nível de significância), o teste é executado novamente, mas agora usando a hipótese nula de ($M_1 \geq M_2$). Este processo foi adotado em todos os experimentos realizados neste capítulo,

para garantir uma melhor interpretação dos resultados. Todos os testes estatísticos foram executados utilizando a ferramenta R[1].

3.2.2 Implementação dos Métodos de Pontuação de Sentenças

Nesta subseção são apresentadas algumas decisões de implementação dos métodos de pontuação das sentenças. As decisões relativas a limiares ou outros parâmetros foram tomadas baseadas em outros trabalhos na literatura e também em um experimento anterior que avaliou o impacto de diferentes parâmetros.

- **Centralidade das Sentenças**: Três versões das medidas de centralidade de sentenças foram implementadas: **(i)** *Centralidade Sobreposição* que computa a similaridade entre duas sentenças baseada na intersecção de palavras entre elas; **(ii)** *Centralidade Cosseno* que usa a similaridade do cosseno para calcular a semelhança entre as sentenças; e **(iii)** *Centralidade BLEU* (HAQUE et al., 2010) que utiliza o algoritmo de similaridade *Bilingual Evaluation Understudy* (BLEU) para mensurar a similaridade entre duas sentenças. Para o desenvolvimento do algoritmo de similaridade BLEU, utilizou-se a implementação disponibilizada em HultigLib[2].
- **Entidades Nomeadas**: Além das categorias tradicionais como, nome de pessoas, cidades e organizações, algumas ferramentas de PLN, como o Stanford CoreNLP também adotam outras classificações, como datas e porcentagens. Por isso, essas categorias também foram consideradas como entidades durante o desenvolvimento desta técnica.
- **Expressões Chave**: Para implementar este método, a lista de expressões definidas em [3] foi utilizada.
- **Frequência das Palavras:** Como mencionado na Subseção 3.1.1 este método pode levar em consideração todas as palavras exceto as *stopwords*, ou então levar em consideração apenas as N palavras mais frequentes. Baseado na boa performance apresentada pelo sistema Classifier4J (LOTHIAN, 2003) em uma avaliação prévia (BATISTA et al., 2015), decidiu-se definir $N = 100$.
- **Posição das Sentenças**: Duas versões deste método foram implementadas com base nas estratégias apresentadas na Subseção 3.1.1. *Posição das Sentenças versão 1* usa a estratégia adotada em Ferreira et al. (2013) de atribuir maior peso para sentenças no início e no fim dos documentos. Por outro lado, *Posição das Sentenças versão 2* atribui maior peso apenas para sentenças no início dos documentos.
- **Relações Abertas**: Para computar este método é necessário realizar a extração das relações abertas presentes no(s) documento(s) de entrada. Para isso, utilizou-se

[1] http://www.r-project.org/
[2] http://www.di.ubi.pt/ jpaulo/hultiglib/
[3] http://www.cs.otago.ac.nz/staffpriv/alik/papers/apps.pdf

o sistema de extração de relações abertas *Reverb* (FADER; SODERLAND; ETZIONI, 2011).

- **Similaridade Agregada** e **Bushy Pathy**: Ambos os algoritmos foram implementados usando a similaridade do cosseno e adotando um limiar de similaridade de 0,1.
- **Similaridade Léxica**: O algoritmo de similaridade semântica *Resnik* (RESNIK, 1995), que usa as informações presentes no *WordNet* [4], foi utilizado para calcular a similaridade entre duas palavras.
- **Tamanho das Sentenças**: Antes de computar a pontuação das sentenças usando este método, sentenças com menos de 10 palavras ou com mais de 50 palavras foram removidas. *Stopwords* não foram levadas em consideração durante a contagem. As sentenças não removidas recebem sua pontuação baseada na Equação 3.16.
- **TextRank**: Este método foi implementado seguindo a sugestão de Barrera e Verma (2012), que identificou melhores resultados utilizando um modelo baseado em 4-gramas e levando em consideração apenas substantivos e adjetivos.

Uma importante questão que precisa ser definida são critérios de escolha quando duas ou mais sentenças possuem a mesma pontuação, principalmente para a sumarização multidocumento. Para solucionar esse problema os seguintes critérios foram adotados:

- **Posição das Sentenças:** Quando duas ou mais sentenças possuem a mesma pontuação nesse método, utilizou-se a técnica de frequência de palavras como critério de escolha. Por exemplo, imagine que duas sentenças s_1 e s_2 possuem a mesma pontuação e somente uma pode ser inserida no resumo gerado por conta do limiar do tamanho máximo do resumo. Nesse cenário, a sentença com maior escore na técnica de frequência de palavras foi selecionada para o resumo gerado.
- **Demais técnicas:** Para todas as demais técnicas investigadas, aplicou-se o método de posição das sentenças como critério de desempate.

3.2.3 Avaliação Individual dos Métodos de Pontuação de Sentenças

Neste experimento, cada técnica de pontuação de sentenças apresentada na Seção 3.1 é avaliada individualmente nas tarefas de sumarização monodocumento e multidocumento. O objetivo deste experimento é identificar quais as melhores técnicas de pontuação de sentenças para cada tarefa de sumarização (monodocumento e multidocumento).

[4] http://wordnet.princeton.edu/

Sumarização Monodocumento

Na Tabela 2 são apresentados os resultados da avaliação individual dos métodos de pontuação das sentenças com base nas medidas de cobertura do ROUGE-1 (R-1) e ROUGE-2 (R-2), além da medida de Intersecção de Sentenças (IS) entre os resumos gerados e os resumos de referência. Os dez métodos com melhor performance em cada corpus são destacados em negrito. Variações do mesmo método de pontuação não foram selecionados duas vezes para evitar redundâncias, ou seja, caso o método de Posição de Sentença versão 1 e Posição de Sentença versão 2 estejam entre os dez melhores métodos, apenas o método com melhor performance é selecionado. A mesma ideia é aplicada para os métodos de Centralidade das Sentenças, Centralidade das Sentenças Cosseno e Centralidade das Sentenças BLEU. Os resultados obtidos variam substancialmente de uma técnica para outra, demonstrando que uma grande diversidade de resumos é gerada pelos métodos.

No corpus CNN, o método FT-FIS obteve o melhor resultado baseado nas medidas do R-1 e R-2, apresentando uma diferença significativa ao nível de 95% de confiança em ambas as medidas comparado com os resultados dos demais métodos. Com relação à medida de IS, o método de similaridade com o título apresentou o melhor resultado, identificando 26,91% das 10.754 sentenças presentes nos resumos de referência. É possível observar que existe uma alta correlação entre as técnicas selecionadas pelas três medidas de avaliação (R-1, R-2 e IS), reforçando ainda mais os bons resultados das técnicas selecionadas.

Nos experimentos realizados no DUC 2001, o método de posição das sentenças versão 2 apresentou os melhores resultados em termos das medidas do R-1 e R-2. Baseado na medida do R-1, a técnica de posição das sentenças versão 2 apresentou resultados estatisticamente superiores a todos os outros métodos, com exceção do método de similaridade com o título. Em relação à medida do R-2, a técnica de posição das sentenças versão 2 supera significativamente todas as demais técnicas.

No corpus do DUC 2002, mais uma vez o método de posição das sentenças versão 2 apresentou resultados estatisticamente superiores nas medidas do R-1 e R-2, em relação a todos os outros métodos. O método de posição das sentenças tem demonstrado uma alta performance nos corpora do DUC. Vale ressaltar que esse método foi usado como *baseline* nas competições do DUC 2001 e 2002 na tarefa de sumarização monodocumento.

Em todos os três corpora, o método de posição das sentenças versão 2 apresentou melhores resultados do que a técnica de posição das sentenças versão 1. Isso corrobora com o fato de que as primeiras sentenças em artigos de notícias são usualmente as mais importantes e, portanto, possuem uma alta probabilidade de serem incluídas no resumo. Isso acontece devido ao estilo de escrita de textos de notícias que, em geral, introduzem os principais fatos da notícia no início do documento e nas sentenças seguintes descrevem tais fatos em mais detalhes. O método de Posição das Sentenças versão 1, que também atribui maior peso para sentenças no início e no fim do documento, é mais apropriado para documentos grandes tais como livros, artigos científicos, entre outros.

Tabela 2 – Resultados (%) e desvio padrão entre parênteses da avaliação dos métodos de pontuação de sentenças na tarefa de sumarização monodocumento. Os dez métodos com melhor performance em cada corpus são destacados em negrito. O método com melhor performance é indicado por * e o grupo de métodos estatisticamente similar a ele, se existir, é indicado usando †.

Técnicas	**CNN**			**DUC 01**		**DUC 02**	
	IS	R-1	R-2	R-1	R-2	R-1	R-2
Bushy Path	**19,04**	41,11 (18,23)	24,28 (20,74)	**42,28** (9,69)	**16,31** (10,82)	**44,58** (9,25)	**18,90** (10,02)
Centr. BLEU	6,59	22,18 (18,82)	13,58 (19,38)	36,28 (10,44)	11,88 (9,63)	39,42 (9,66)	14,50 (9,60)
Centr. Cosseno	18,46	40,73 (18,49)	23,99 (20,91)	**41,75** (9,10)	**15,79** (9,84)	**44,25** (9,25)	**18,72** (9,99)
Centr. Sobreposição	**21,31**	**53,34** (19,62)	**32,75** (24,87)	40,26 (10,27)	14,86 (10,67)	43,70 (9,86)	18,18 (10,42)
Coocorrência das Palavras	13,28	41,74 (20,27)	22,71 (22,85)	37,17 (9,33)	12,20 (8,97)	38,85 (9,57)	13,80 (9,26)
Dados Numéricos	12,80)	31,99 (18,11)	17,74 (18,74)	35,65 (11,96)	12,68 (10,02)	37,25 (10,84)	14,21 (9,04)
Entidades Nomeadas	**22,24**	**48,92** (19,88)	**31,33** (23,98)	**39,42** (10,40)	**15,11** (10,17)	**42,93** (10,21)	**18,07** (10,17)
Expressões Chave	15,41	41,28 (18,58)	22,33 (21,26)	38,50 (9,40)	13,79 (9,60)	40,35 (9,67)	15,65 (9,10)
Freq. das Palavras	**24,34**	**52,40** (19,80)	**34,33** (24,90)	**41,83** (10,37)	**16,54** (11,01)	**44,77** (9,42)	**19,59** (9,96)
FT-FIS	**23,99**	**54,25*** (19,87)	**35,65*** (25,44)	**40,73** (10,46)	**16,01** (10,94)	**44,00** (9,72)	**18,70** (10,06)
Nomes Próprios	13,11	31,90 (17,10)	17,38 (18,05)	37,21 (10,05)	13,45 (9,55)	40,20 (10,34)	15,95 (9,97)
Palavras Maiúsculas	12,04	29,54 (16,92)	16,31 (17,43)	37,47 (10,17)	13,65 (9,78)	40,57 (10,58)	16,36 (10,15)
Pos. Sent. versão 1	18,17	39,00 (19,20)	24,68 (21,73)	41,49 (10,41)	17,61 (11,05)	44,38 (9,08)	19,72 (9,67)
Pos. Sent. versão 2	**26,74**	**45,99** (21,77)	**33,49** (25,00)	**43,75*** (10,47)	**19,57*** (11,64)	**46,94*** (9,20)	**22,14*** (10,01)
Relações Abertas	**17,76**	**46,42** (19,54)	**26,97** (23,73)	**39,12** (10,27)	**14,41** (10,78)	**43,01** (9,77)	**17,57** (10,39)
Sim. com o Título	**26,91***	**49,29** (20,50)	**34,51** (23,67)	**42,59†** (10,20)	**17,93** (10,62)	**44,31** (10,74)	**19,95** (10,24)
Sim. Agregada	16,53	36,82 (17,64)	20,98 (19,29)	**41,16** (9,54)	**15,27** (10,10)	**43,73** (9,54)	**18,26** (10,08)
Sim. Léxica	16,92	45,62 (18,82)	25,65 (22,85)	38,59 (9,97)	13,59 (10,01)	41,27 (10,20)	16,09 (10,32)
Sintagmas Nominais e Verbais	18,78	**50,70** (19,59)	**30,05** (24,78)	38,68 (10,39)	13,60 (9,80)	41,47 (10,08)	16,38 (10,04)
Tam. das Sentenças	**19,32**	**52,67** (19,87)	**31,69** (25,75)	38,31 (10,38)	13,43 (10,19)	41,51 (9,85)	16,27 (9,98)
TextRank	**21,99**	**49,74** (20,03)	**31,59** (24,36)	**40,66** (9,38)	**15,09** (9,76)	**43,93** (9,77)	**18,66** (10,11)

Geralmente, o título de artigos de notícias proporciona uma importante indicação do principal assunto discutido. Os títulos dos documentos do corpus CNN, em geral, são bem escritos e muito descritivos, esse fato levou a técnica de similaridade com o título a obter boa performance nas medidas do R-1 e R-2, e a melhor performance na medida de IS. O título dos documentos nos corpora do DUC não são tão descritivos como nos documentos do CNN, mesmo assim, a similaridade com o título obteve bons resultados nesses corpora.

Outras técnicas como similaridade agregada, *bushy path*, centralidade com sobreposição, centralidade com o cosseno, *textrank*, FT-FIS, e frequência das palavras também apresentaram boa performance. Os métodos de centralidade obtiveram bom desempenho em todos os corpora, demonstrando que sentenças compartilhando muitas informações com outras sentenças são boas candidatas para serem incluídas no resumo. As sentenças contendo mais palavras consideradas importantes também possuem uma alta probabilidade de serem inseridas no resumo, os métodos de FT-FIS, frequência das palavras e *Textrank* apresentaram bons resultados em reconhecer termos importantes nos documentos.

Sumarização Multidocumento

Na Tabela 3 são apresentados os resultados da avaliação dos métodos de pontuação de sentenças na tarefa de sumarização multidocumento. Já que o método de Pontuação de Sentenças versão 2 apresentou melhores resultados do que a Posição de Sentenças versão 1 na avaliação da sumarização monodocumento, somente a versão 2 será avaliada neste experimento. De maneira similar ao experimento anterior, uma grande diversidade de resumos e variação na performance dos métodos com base nas medidas do ROUGE foram observadas.

No DUC 2001, os métodos de similaridade com o título, posição das sentenças versão 2, centralidade do cosseno, *bushy path* e similaridade agregada apresentaram os cinco melhores resultados com base na medida R-1. Não existe diferença estatística entre eles ao nível de 95% de confiança. A técnica de similaridade com o título também obteve a melhor performance baseada na medida R-2, contudo os métodos de centralidade do cosseno, similaridade agregada, *bushy path*, posição das sentenças versão 2 e frequência das palavras obtiveram resultados estatisticamente similares.

Nos experimentos usando o DUC 2002, os métodos de posição das sentenças versão 2 e centralidade do cosseno apresentaram o melhor desempenho com base no R-1 e R-2, respectivamente. Em ambas as medidas, nenhuma diferença estatística foi observada entre eles e os métodos de *Textrank*, *bushy path*, similaridade com o título e similaridade agregada.

No DUC 2003, a técnica de centralidade do cosseno obteve a melhor performance com base no R-1, contudo tal resultado não apresentou diferença estatística em relação às medidas de similaridade agregada, *bushy path*, similaridade com o título, posição das

Tabela 3 – Resultados (%) e desvio padrão entre parênteses da avaliação dos métodos de pontuação de sentenças na tarefa de sumarização multidocumento. Os dez métodos com melhor performance em cada corpus são destacados em negrito. O método com melhor performance é indicado por * e o grupo de métodos estatisticamente similar a ele, se existir, é indicado por †.

Métodos	DUC 01		DUC 02		DUC 03		DUC 04	
	R-1	R-2	R-1	R-2	R-1	R-2	R-1	R-2
Bushy Path	**29,67†** (6,21)	**5,12†** (3,12)	**32,67†** (5,36)	**6,49†** (3,11)	**36,96†** (7,28)	**8,39†** (5,03)	**36,85*** (4,97)	**8,34*** (3,75)
Cent. BLEU	22,35 (5,55)	1,92 (1,14)	24,11 (6,31)	2,76 (2,05)	23,83 (6,10)	3,30 (2,42)	25,88 (5,14)	3,15 (1,85)
Cent. Cosseno	**30,25†** (6,18)	**5,50†** (3,37)	**33,33†** (4,97)	**7,29*** (3,14)	**37,61*** (8,00)	**9,42†** (5,43)	**36,69†** (4,45)	**8,31†** (3,23)
Cent. Sobreposição	26,34 (7,33)	3,62 (4,35)	28,73 (6,42)	4,32 (3,02)	33,75 (6,21)	6,63 (4,50)	30,22 (5,85)	4,69 (2,60)
Coocorrência das Palavras	23,72 (5,66)	2,60 (1,88)	26,40 (6,59)	3,75 (3,04)	**33,18** (7,01)	**6,04** (4,36)	30,06 (5,85)	4,88 (2,69)
Dados Numéricos	21,94 (7,53)	3,08 (2,16)	24,72 (5,53)	3,26 (2,30)	28,78 (7,75)	5,85 (4,32)	26,99 (6,30)	4,06 (3,06)
Entidade Nomeadas	23,90 (6,69)	**3,52** (3,33)	26,83 (7,29)	**4,17** (2,90)	**33,22** (7,24)	**6,46** (3,47)	**30,75** (6,96)	**5,53** (3,24)
Expressões Chave	**25,70** (5,42)	**3,41** (2,95)	**29,01** (5,98)	**4,37** (2,81)	29,95 (5,83)	5,03 (3,44)	28,92 (5,04)	4,06 (2,14)
Freq. das Palavras	**28,28** (7,15)	**4,97†** (3,95)	**30,17** (7,82)	**5,39** (3,28)	**36,68†** (7,66)	**8,41** (5,68)	**35,83†** (4,61)	**8,15†** (3,07)
FT-FIS	**27,79** (8,18)	**4,71** (4,09)	**30,08** (7,85)	**5,58** (3,68)	**35,64†** (7,29)	**7,69** (5,62)	**35,55†** (6,57)	**8,18†** (3,58)
Nomes Próprio	20,42 (6,52)	2,84 (2,54)	23,34 (6,76)	4,15 (2,77)	28,67 (7,21)	5,38 (3,60)	28,52 (5,60)	5,02 (2,79)
Palavras Iniciais Maiúsculas	21,11 (6,52)	3,02 (2,37)	23,32 (6,51)	3,95 (2,54)	29,01 (6,79)	5,73 (3,61)	28,48 (5,72)	**5,31** (2,85)
Pos. das Sent. versão 2	**30,29†** (5,93)	**5,10†** (3,98)	**33,78*** (5,65)	**7,16†** (3,69)	**35,93†** (7,23)	**7,67** (3,98)	**35,78†** (4,52)	**7,87†** (3,17)
Relações Abertas	**25,78** (4,92)	2,99 (1,91)	**28,43** (5,26)	3,84 (2,77)	32,24 (6,41)	5,57 (3,46)	**31,10** (4,10)	4,88 (2,31)
Sim. com o Título	**30,63*** (6,66)	**5,94*** (3,74)	**32,60†** (6,17)	**6,87†** (3,56)	**35,79†** (6,87)	**7,84** (5,23)	**35,45†** (4,92)	**7,91†** (3,20)
Sim. Agregada	**29,21†** (6,99)	**5,33†** (3,36)	**32,58†** (5,03)	**6,71†** (3,23)	**37,59†** (8,37)	**9,55*** (5,27)	**35,87†** (5,64)	**8,30†** (3,41)
Sim. Léxica	24,39 (5,39)	2,64 (1,90)	26,91 (6,52)	3,80 (2,76)	30,76 (6,09)	5,03 (2,57)	29,02 (5,99)	4,66 (2,80)
Sintagmas Nominais e Verbais	**24,67** (6,24)	2,79 (2,64)	26,68 (6,29)	3,35 (2,29)	31,12 (7,14)	5,13 (4,03)	28,16 (5,21)	3,60 (1,73)
Tam. das Sentenças	25,12 (6,00)	2,97 (2,77)	27,01 (6,61)	3,68 (2,79)	31,96 (5,10)	5,51 (3,90)	27,95 (5,29)	3,66 (2,26)
TextRank	**28,72** (6,80)	**5,00** (3,40)	**33,62†** (6,09)	**6,98†** (3,95)	**35,75†** (7,08)	**6,89** (3,85)	**35,85†** (5,19)	**7,73†** (3,26)

sentenças versão 2, *TextRank*, FT-FIS e frequência das palavras. Levando em consideração o R-2, os métodos de similaridade agregada e centralidade do cosseno apresentaram os dois melhores resultados. A técnica de similaridade agregada apresentou uma melhora significativa em relação a todos os outros métodos, com exceção do método de centralidade do cosseno.

Na avaliação usando o DUC 2004, o método de *bushy path* apresentou o melhor resultado com base no R-1 e R-2. Contudo, os métodos de centralidade do cosseno, similaridade agregada, *textrank*, frequência das palavras, posição das sentenças versão 2, FT-FIS e similaridade com o título obtiveram resultados estatisticamente similares em ambos as medidas do ROUGE.

As técnicas de similaridade agregada, *bushy path*, centralidade do cosseno e posição de sentenças versão 2 obtiveram o melhor desempenho na tarefa de sumarização multidocumento levando em consideração os quatro corpora utilizados. Este resultado demonstra que a suposição de que sentenças que compartilham muitas informações com outras sentenças, podem ser consideradas as mais importantes. Tal suposição ainda é mais forte na sumarização multidocumento dada a alta redundância de informações entre os documentos do mesmo grupo. A técnica de posição das sentenças não foi tão efetiva como na sumarização monodocumento, mas ainda assim obteve bons resultados. A técnica de similaridade com o título também apresentou uma boa performance, reforçando a importância do título para a identificação do tema principal abordado nos documentos. As técnicas de frequência das palavras, *Textrank* e FT-FIS também obtiveram uma boa performance.

Os experimentos realizados nas duas tarefas possibilitaram observar que os métodos de pontuação de sentenças apresentam resultados diferentes dependendo da tarefa em questão e do corpus utilizado. Levando em consideração todos os corpora utilizados em ambas as tarefas de sumarização, é possível identificar que os métodos de *bushy path*, centralidade do cosseno, centralidade com sobreposição, frequência das palavras, posição das sentenças versão 2, similaridade agregada, similaridade com o título, *Textrank* e FT-FIS, apresentaram globalmente os melhores resultados. Tais métodos foram considerados neste trabalho como os mais adequados para a tarefa de pontuação das sentenças.

3.2.4 Avaliando Estratégias de Combinação

Nesta seção são apresentados os experimentos conduzidos para avaliar quatro estratégias para combinação das medidas de pontuação de sentenças investigadas na Subseção 3.2.3. Para isso, foram selecionados os dez métodos com melhor performance global nos experimentos descritos na seção anterior com base na medida de cobertura do ROUGE-1 (R-1). A medida do R-1 foi adotada devido aos bons resultados que trabalhos recentes (SIPOS; SHIVASWAMY; JOACHIMS, 2012; HONG; MARCUS; NENKOVA, 2015) obtiveram utilizando-a como critério de seleção entre diferentes métodos de sumarização.

Uma grande variedade de resumos foi gerada usando cada método de pontuação de sentenças individualmente. Baseado neste fato, é razoável presumir que algumas dessas técnicas podem ser complementares umas com as outras. Dessa forma, o objetivo deste experimento é identificar as melhores combinações dos métodos de pontuação de sentenças, buscando obter resultados superiores aos obtidos utilizando as técnicas individualmente.

As quatro estratégias de combinação a seguir foram analisadas:

- **Combinação Linear**: Nesta combinação, a pontuação final de uma sentença é dada pela média aritmética da pontuação individual de cada um dos **N** métodos de pontuação de sentenças em questão.
- **Combinação Linear Ponderada**: Esta combinação é uma extensão da estratégia anterior, só que considerando pesos diferentes para cada método de pontuação adotado. Identificar a melhor configuração de pesos para cada método é um problema clássico de otimização. Para solucionar esse problema, Algoritmos Genéticos (GOLDBERG, 1989) foram utilizados. Baseado em trabalhos anteriores em SAT (ABUOBIEDA et al., 2012; ABUOBIEDA et al., 2013), definiu-se o tamanho da população como 50 e o número máximo de geração igual a 100. A medida R-1 foi adotada como função de *fitness* e três operadores evolucionários para gerar novas populações foram usados: Seleção, *Crossover* e Mutação. Como método de avaliação utilizou-se a Validação Cruzada com **k**-*folders*. Na sumarização monodocumento definiu-se $k = 10$, enquanto que na sumarização multidocumento k é igual ao número de grupos que o corpus possui. Dessa forma, **k - 1** *folders* são usados como conjunto de treinamento para identificação da melhor configuração de pesos, e o *folder* não selecionado para treinamento é usado como conjunto de teste. Esse processo é repetido até que todos os *folders* sejam selecionados como conjunto de teste.
- **Voto Majoritário**: Neste método de combinação, o conjunto de M métodos de pontuação de sentenças são considerados *Eleitores*. Cada método $m \in M$ seleciona diferentes sentenças para compor o resumo, e cada uma dessas sentenças recebe um voto. Após todos os métodos serem executados, as sentenças com maior número de votos são selecionadas para compor o resumo final. No caso de empate entre duas ou mais sentenças, aquela mais próxima do início do documento é privilegiada.
- ***Condorcet Ranking*** **(PALSHIKAR; DESHPANDE; ATHIAPPAN, 2012)**: Esta estratégia de combinação é similar à anterior, a diferença é que a posição da sentença no ranking de pontuação de cada método $m \in M$ adotado é levada em consideração. Para isso, essa estratégia de combinação computa os votos de cada método $m \in M$ para decidir as sentenças vencedoras entre o conjunto S de sentenças. Uma matriz K SxS é criada e usada para comparar cada sentença s_i (linha em K) contra as outras sentenças s_j (coluna em K). Se uma sentença s_i é selecionada por um método m com uma pontuação maior que uma outra sentença s_j, então s_i recebe um voto e

a entrada da matriz $K[i, j]$ é atualizada. Após o processo de votação, as sentenças com maior número de votos vencedores são selecionadas e inseridas no resumo.

Todas as possíveis combinações dos dez melhores métodos de pontuação de sentenças selecionados com base na medida R-1 nos experimentos descritos na Subseção 3.2.3 foram analisadas. Dado que existem 10 métodos de pontuação em consideração, e não sendo permitida a repetição desses métodos, é possível gerar 1.024 combinações. Removendo as combinações que possuem apenas um método e a combinação nula, no total foram analisadas 1.013 combinações. Cada combinação gerada é avaliada utilizando as estratégias de combinação Linear, Voto Majoritário e *Condorcet Ranking*. Dado o esforço computacional necessário para identificar a melhor configuração de pesos usando Algoritmos Genéticos, essa estratégia só foi adotada nas duas melhores combinações lineares identificadas com base na medida do ROUGE-1 em cada corpus utilizado.

Para facilitar a apresentação das melhores combinações, no Quadro 5 são apresentados rótulos para os nomes das técnicas de pontuação de sentenças usadas.

Quadro 5 – Rótulos das técnicas de pontuação de sentenças usadas nas avaliações das estratégias de combinação.

Técnicas	**Rótulos**
Bushy Path	BusPath
Centralidade Cosseno	CentCos
Centralidade Sobreposição	CentSob
Entidade Nomeadas	EntNom
Expressões Chave	ExpChave
Frequência das Palavras	FreqPal
FT-FIS	FT-FIS
Posição das Sentenças versão 2	PosSentV2
Relações Abertas	RelAb
Similaridade com o Título	SimTit
Similaridade Agregada	SimAg
Sintagmas Nominais e Verbais	NP_VP
Tamanho das Sentenças	TamSent
TextRank	TextRank

Sumarização Monodocumento

No Quadro 6 são apresentadas as duas combinações que obtiveram os melhores resultados com base na medida de cobertura do R-1, para cada estratégia de combinação analisada. Na Tabela 4 são apresentados os resultados dessas combinações em termos das

medidas de cobertura do R-1, R-2 e da medida de IS. É possível observar que as melhores combinações obtiveram resultados similares em todos os corpora.

Quadro 6 – Lista das duas melhores combinações com base na medida ROUGE-1 para cada estratégia de combinação na sumarização monodocumento.

Combinação	**CNN**
Condorcet A	EntNom, NP_VP, RelAb, CentSob, TamSent, PosSentV2, SimTit, TextRank, FreqPal
Condorcet B	EntNom, NP_VP, RelAb, CentSob, TamSent, PosSentV2, SimTit, TextRank, FT-FIS
Combinação Linear A	CentSob, PosSentV2, FT-FIS
Combinação Linear B	CentSob, TamSent, PosSentV2, FreqPal
Voto Majoritário A	EntNom, NP_VP, CentSob, TamSent, PosSentV2, SimTit, TextRank, FreqPal
Voto Majoritário B	EntNom, RelAb, CentSob, TamSent, PosSentV2, SimTit, TextRank, FT-FIS, FreqPal
Combinação	**DUC 2001**
Condorcet A	PosSentV2, FreqPal
Condorcet B	BusPath, RelAb, PosSentV2, SimTit
Combinação Linear A	CentCos, PosSentV2
Combinação Linear B	BusPath, PosSentV2
Voto Majoritário A	BusPath, PosSentV2, SimTit
Voto Majoritário B	BusPath, PosSentV2, FreqPal
Combinação	**DUC 2002**
Condorcet A	PosSentV2, FreqPal
Condorcet B	PosSentV2, FT-FIS
Combinação Linear A	BusPath, PosSentV2, FT-FIS
Combinação Linear B	SimAg, PosSentV2, FT-FIS
Voto Majoritário A	BusPath, PosSentV2, SimTit
Voto Majoritário B	SimAg, PosSentV2, SimTit

Nos experimentos realizados no corpus CNN, a combinação linear A e a combinação linear ponderada A obtiveram a melhor performance com base no R-1 e R-2, respectivamente. Tais combinações superaram estatisticamente as estratégias de condorcet e voto majoritário, mas obtiveram resultados estatisticamente similares com a combinação linear B e combinação linear ponderada B. A combinação linear ponderada A também apresentou o melhor resultado em termos da medida IS.

Em relação ao DUC 2001, a combinação linear ponderada A apresentou os melhores resultados analisando as medidas do R-1 e R-2. Essa combinação obteve resultados estatisticamente superiores em relação as combinações condorcet A e B, e voto majoritário B com base no R-1. Já em termos da medida R-2, a combinação linear ponderada A apresentou melhorias significativas em relação as combinações condorcet A e B, e a combinação Linear B.

No DUC 2002, a melhor performance com base no R-1 e R-2 foi obtida pela combinação linear ponderada A e a combinação linear ponderada B, respectivamente. Baseado no R-1, a combinação linear ponderada A obteve resultados estatisticamente superiores em relação à todas as combinações usando condorcet e voto majoritário. Já com base no R-2, a combinação linear ponderada B somente apresentou uma melhora significativa em

Tabela 4 – Resultados (%) e desvio padrão entre parênteses das duas melhores combinações para estratégia de agregação na sumarização monodocumento. A combinação com melhor performance em cada corpus é destacada em negrito e o grupo de combinações estatisticamente similar, se existir, é indicado usando †.

Combinações	**CNN**		
	IS	R-1	R-2
Condorcet A	27,47	57,00 (20,37)	39,40 (25,80)
Condorcet B	26,77	56,86 (20,20)	38,91 (25,69)
Combinação Linear A	29,89	**57,93** (20,22)	41,54† (25,32)
Combinação Linear B	28,39	57,87† (20,18)	40,60† (25,63)
Voto Majoritário A	27,76	56,98 (20,42)	39,63 (25,75)
Voto Majoritário B	27,92	56,95 (20,27)	39,63 (25,55)
Comb. Linear Ponderada A	**30,05**	57,86† (20,20)	**41,68** (25,17)
Comb. Linear Ponderada B	29,53	57,89† (20,34)	41,53† (25,41)
Combinações	**DUC 2001**		
	IS	R-1	R-2
Condorcet A	-	43,70 (10,32)	19,50 (11,60)
Condorcet B	-	44,06 (9,78)	19,40 (11,20)
Combinação Linear A	-	44,82† (9,53)	19,72† (11,09)
Combinação Linear B	-	44,57† (9,74)	19,40 (11,20)
Voto Majoritário A	-	44,28† (10,08)	19,72† (11,50)
Voto Majoritário B	-	44,30 (10,48)	19,78† (11,88)
Comb. Linear Ponderada A	-	**44,98** (9,56)	**20,12** (11,04)
Comb. Linear Ponderada B	-	44,72† (9,51)	19,72† (10,93)
Combinações	**DUC 2002**		
	IS	R-1	R-2
Condorcet A	-	46,66 (9,20)	21,89 (9,87)
Condorcet B	-	46,63 (9,28)	21,90 (9,96)
Combinação Linear A	-	47,55† (8,60)	22,06† (9,74)
Combinação Linear B	-	47,53† (8,70)	22,13† (9,87)
Voto Majoritário A	-	47,22 (9,38)	22,17† (10,32)
Voto Majoritário B	-	47,19 (9,25)	22,14† (10,22)
Comb. Linear Ponderada A	-	**47,73** (8,47)	22,34† (9,56)
Comb. Linear Ponderada B	-	47,64† (8,59)	**22,39** (9,74)

comparação com as combinações condorcet A e B.

Sumarização Multidocumento

No Quadro 7 são apresentadas as duas melhores combinações encontradas nos experimentos realizados com base na medida de cobertura do ROUGE-1 para cada um dos métodos de combinação investigados. Na Tabela 5 são resumidos os resultados dessas combinações em termos das medidas de cobertura do R-1, R-2 e da medida de IS. Assim como na sumarização monodocumento, neste experimento, as combinações apresentaram

resultados similares, principalmente na medida do R-2.

Quadro 7 – Melhores combinações com base na medida ROUGE-1 para cada estratégia de agregação na sumarização multidocumento.

Combinações	**DUC 2001**	**DUC 2002**
Condorcet A	CentCos, PosSentV2	CentCos, PosSentV2, TextRank
Condorcet B	PosSentV2, TextRank	SimAg, PosSentV2
Combinação Linear A	SimAg, ExpChave, PosSentV2, TextRank	SimAg, BusPath, CentCos, PosSentV2, SimTit, FreqPal
Combinação Linear B	BusPath, PosSentV2, TextRank	SimAg, PosSentV2, TextRank
Voto Majoritário A	ExpChave, CentCos, PosSentV2	BusPath, PosSentV2
Voto Majoritário B	SimAg, RelAb, PosSentV2	CentCos, PosSentV2
Combinações	**DUC 2003**	**DUC 2004**
Condorcet A	SimAg, BusPath, PosSentV2, TextRank	SimAg, RelAb, CentCos, PosSentV2, FT-FIS
Condorcet B	SimAg, CentCos, PosSentV2, TextRank	BusPath, RelAb, CentCos, PosSentV2, FT-FIS
Combinação Linear A	SimAg, CentCos, TextRank	SimAg, BusPath, EntNom, PosSentV2, TextRank
Combinação Linear B	SimAg, TextRank	SimAg, EntNom, CentCos, PosSentV2, TextRank
Voto Majoritário A	SimAg, BusPath, EntNom, PosSentV2, TextRank	BusPath, RelAb, CentCos, PosSentV2, SimTit, FT-FIS
Voto Majoritário B	SimAg, BusPath, FreqPal	CentCos, PosSentV2, TextRank

Tabela 5 – Resultados (%) e desvio padrão entre parênteses das duas melhores combinações para estratégia de agregação na sumarização multidocumento. A combinação com melhor performance em cada corpus é destacada em negrito e o grupo de combinações estatisticamente similar, se existir, é indicado usando †.

Combinações	**DUC 2001**		**DUC 2002**	
	R-1	R-2	R-1	R-2
Condorcet A	32,29† (6,42)	7,12† (5,77)	35,67† (5,20)	**8,04** (3,56)
Condorcet B	32,32† (7,11)	6,98† (6,00)	33,82 (5,04)	7,27 (3,42)
Combinação Linear A	32,45 (8,20)	7,51† (6,04)	**35,83** (5,06)	7,83† (3,59)
Combinação Linear B	32,55 (7,14)	6,91† (4,87)	34,99† (5,06)	8,02† (3,93)
Voto Majoritário A	31,89 (7,07)	7,00† (6,18)	33,87 (5,04)	7,27 (3,42)
Voto Majoritário B	31,55 (6,43)	7,16† (5,99)	5,70† (4,97)	**8,04** (3,40)
Comb. Linear Ponderada A	**33,63** (7,79)	**7,67** (6,10)	35,55† (5,34)	7,80† (3,58)
Comb. Linear Ponderada B	32,96† (7,84)	7,45† (6,04)	34,73 (5,07)	7,93† (3,45)
Combinações	**DUC 2003**		**DUC 2004**	
	R-1	R-2	R-1	R-2
Condorcet A	39,20† (7,12)	8,96† (4,52)	38,06† (4,88)	9,34† (3,23)
Condorcet B	38,41 (6,84)	9,42† (4,71)	38,24† (4,49)	9,46† (3,25)
Combinação Linear A	39,10† (7,98)	9,28† (5,18)	**38,58** (4,23)	**9,80** (2,98)
Combinação Linear B	38,68† (7,74)	8,94† (4,71)	38,50† (4,18)	9,70† (3,04)
Voto Majoritário A	38,84† (7,31)	8,86 (4,97)	38,37† (4,27)	9,43†(2,91)
Voto Majoritário B	38,40 (6,84)	8,88 (4,85)	37,81† (4,48)	9,48† (3,18)
Comb. Linear Ponderada A	38,52 (9,17)	9,43† (5,84)	*38,15*† (4,29)	9,58† (3,02)
Comb. Linear Ponderada B	**39,61** (9,14)	**9,76** (5,39)	37,68 (4,33)	9,52† (2,82)

A combinação linear ponderada A obteve a melhor performance com base nas medidas do R-1 e R-2 no DUC 2001. Os resultados obtidos por essa combinação foram estatisticamente superiores em comparação com as combinações lineares A e B, e as combinações usando voto majoritário em termos do R-1. Com base no R-2, os resultados de todas as combinações foram muito próximos, dessa forma, todas as combinações apresentaram resultados estatisticamente similares.

No corpus do DUC 2002, a combinação linear A apresentou o melhor resultado em termos do R-1. Tal resultado foi significativamente superior em relação às combinações *condorcet* B, voto majoritário A, e combinação linear ponderada B. Com base no R-2, as combinações *condorcet* A e voto majoritário B obtiveram a melhor performance, apresentando resultados superiores as combinações *condorcet* B e voto majoritário A.

Com relação aos resultados obtidos no corpus do DUC 2003, a combinação linear ponderada B apresentou o melhor desempenho nas medidas do R1 e R2. Tal combinação superou estatisticamente os resultados das combinações *condorcet* B, voto majoritário B, e combinação linear ponderada A com relação ao R-1. Já em termos do R-2, a combinação linear ponderada B apresentou melhorias significativas somente quando comparada com as combinações com voto majoritário A e B.

No corpus do DUC 2004, a combinação linear A apresentou a melhor performance com base no R-1 e R-2. Contudo, tais resultados só foram estatisticamente superiores em relação à combinação linear ponderada B em termos do R-1. Todas as combinações obtiveram resultados do R-2 muito próximos, dessa forma, nenhuma diferença estatística entre elas foi observada.

Analisando os resultados dos experimentos na sumarização monodocumento e multidocumento é possível observar que a estratégia de combinação linear apresentou melhor desempenho do que as estratégias condorcet e voto majoritário em quase todas as comparações. Além disso, observando o Quadro 6 e o Quadro 7 é possível ver que a estratégia de combinação linear necessitou de menos métodos de pontuação para obter os melhores desempenhos. A ideia de considerar pesos diferentes utilizando algoritmos genéticos para cada um dos métodos de pontuação nas duas combinações lineares não foi tão efetiva quanto se esperava. Embora essa ponderação tenha melhorado a performance em alguns casos, apenas no corpus DUC 2001 no contexto multidocumento, observou-se uma diferença estatisticamente significativa em relação às combinações lineares simples. Vale ressaltar que essa estratégia não foi eficiente para as combinações investigadas, mas para outras combinações, resultados melhores podem ser encontrados.

Uma grande variedade de combinações apresenta os melhores desempenhos, dependendo do corpus considerado. Como pode ser observado, nenhuma combinação ficou entre as duas melhores performances em mais de um corpus. Esse fato sugere que as características do corpus e a tarefa de sumarização em questão (monodocumento ou multidocumento) possuem um impacto direto na performance dos métodos de pontuação de sentenças. De

maneira geral, nenhum padrão nas combinações das técnicas foi observado nos experimentos realizados, no sentido de indicar que sempre a mesma combinação de técnicas apresenta um alto desempenho para todos o(s) documento(s). Esse comportamento também explica o porquê das combinações ponderadas, apesar de serem supervisionadas, não terem obtido uma melhoria significativa em relação à sua versão não ponderada na maioria dos cenários avaliados. As melhores técnicas nos documentos de treinamento recebem maior peso, mas, em geral, essas técnicas nem sempre também apresentaram bons resultados nos documentos de teste.

Embora diferentes combinações tenham obtido o melhor desempenho em cada corpus, pode-se observar as técnicas que são mais adotadas e suas associações recorrentes. Por exemplo, as seis técnicas mais adotadas nas combinações são: similaridade agregada, *bushy path*, centralidade cosseno, posição das sentenças versão 2, similaridade com o título, e *Textrank*. Na tarefa de sumarização monodocumento, a associação entre os métodos de posição das sentenças versão 2 e similaridade com o título foi a mais recorrente. Já na sumarização multidocumento, a associação entre os métodos de posição das sentenças versão 2, *Textrank*, e uma das medidas de centralidade (similaridade agregada, *bushy path*, ou centralidade cosseno) foi a mais recorrente.

3.2.5 Avaliando Algoritmos de Aprendizagem de Máquina para a Classificação e Pontuação das Sentenças

Este experimento avalia o desempenho dos métodos de pontuação de sentenças apresentados na Seção 3.1 como características para vários algoritmos de aprendizagem de máquina. Dez algoritmos disponíveis na plataforma Weka (MANNING et al., 2014) foram avaliados, sendo eles: *AdaBoostM1* (FREUND; SCHAPIRE, 1996), *J48* (QUINLAN, 1993), *K-nearest Neighbours* (AHA; KIBLER, 1991) referenciado como *IBK*, *Multilayer Perceptron* (HAYKIN, 1998), *Multinomial Logistic Regression* (Logistic) (CESSIE; HOUWELINGEN, 1992), *Naive Bayes* (JOHN; LANGLEY, 1995), *Random Forest* (BREIMAN, 2001), *Random Tree* (QUINLAN, 1992; BREIMAN, 2001), *Radial Basis Function Network* (*RBFNetwork*) (BROOMHEAD; LOWE, 1988), e *Support Vector Machines using Sequential Minimal Optimization* (*SMO*) (PLATT, 1998).

Esses algoritmos foram escolhidos devido à sua popularidade e por representarem diferentes abordagens. Os objetivos deste experimento são: **(i)** Identificar quais algoritmos apresentam melhor desempenho para a classificação das sentenças importantes; e **(ii)** Encontrar o conjunto de características (métodos de pontuação) mais relevantes com base no conjunto de treinamento, aplicando algoritmos de seleção de características. Para isso cada algoritmo é executado usando **(1)** todos os métodos de pontuação de sentenças como características; e **(2)** reduzindo o número de características, aplicando o tradicional algoritmo de *Correlation-based Feature Subset Selection* (CFS) (HALL, 1998), para identificar

as características mais relevantes para o modelo de classificação gerado.

Neste experimento, a tarefa de pontuação das sentenças é tratada como um problema de classificação, no qual algoritmos de AM são utilizados para classificar se uma sentença deve ou não ser incluída no resumo. Um importante aspecto a ser tratado é desbalanceamento no conjunto de treinamento. Tal problema ocorre porque existem mais exemplos de sentenças que não devem ser incluídas nos resumos (classe Não-Resumo) do que exemplos de sentenças que devem ser incluídas nos resumos (classe Resumo). Para contornar esse problema utilizou-se o método *Synthetic Minority Over-sampling TEchnique* (SMOTE) (CHAWLA et al., 2002) para gerar exemplos sintéticos da classe minoritária. O método SMOTE identifica os k vizinhos mais próximos de cada exemplo da classe minoritária, neste trabalho da classe Resumo e, então, cria novos exemplos sintéticos usando a interpolação ao longo do segmento de reta que liga cada exemplo da classe minoritária aos seus vizinhos mais próximos.

Como metodologia de avaliação, o método de validação cruzada (*k-fold cross validation*) foi utilizado. Para cada um dos **k** subconjuntos, as seguintes etapas são realizadas.

- **Treinamento:** Nesta etapa, $k-1$ subconjuntos são usados para treinamento, gerando ao final um modelo de classificação. Na tarefa de sumarização monodocumento o conjunto de documentos foi dividido em dez subconjuntos ($k = 10$), enquanto que na sumarização multidocumento, cada grupo de documentos representa um conjunto, isto é, k é igual ao total de grupos de documentos.

- **Teste:** O conjunto de documentos não selecionado na etapa de treinamento é usado nesta etapa para testar o modelo gerado na etapa anterior. Esse modelo é usado para classificar as sentenças dos documentos de teste nas classes Resumo e Não-Resumo. Para cada classificação o algoritmo atribui uma confiança, que por sua vez é utilizada para ranquear as sentenças classificadas na classe Resumo. Dessa forma, apenas sentenças classificadas como pertencentes a classe Resumo e que possuem uma alta taxa de confiabilidade são selecionadas para compor o resumo final do(s) documento(s).

Os resumos de referência do corpus CNN são extrativos, assim, eles podem ser utilizados diretamente como exemplos durante a fase de treinamento dos algoritmos de classificação. Já os modelos disponíveis para os corpora do DUC são abstrativos. Para resolver esse problema, cada sentença dos resumos de referência do DUC foi mapeada para a sentença com maior similaridade do cosseno no documento original. Dessa forma, as sentenças mapeadas são utilizadas como exemplos da classe Resumo, enquanto que as sentenças não mapeadas são usadas como exemplos da classe Não-Resumo.

Sumarização Monodocumento

Na Tabela 6 são apresentados os resultados dos algoritmos de AM na sumarização monodocumento. A variação de cada algoritmo utilizando o método de seleção de características é indicado na tabela pelo sufixo (CFS).

No corpus CNN, a melhor performance no R-1 foi obtida pelo algoritmo *RBFNetwork*. Contudo, seu desempenho pode ser considerado estatisticamente similar ao nível de 95% de confiança com os algoritmos *Naive Bayes*, *RBFNetwork_CFS*, *Logistic_CFS* e *Logistic*. Em termos do R-2, os dois melhores resultados foram obtidos pelos algoritmos *Logistic_CFS* e *Multilayer Perceptron_CFS*. Nenhuma diferença estatística é observada entre eles e os algoritmos *Naive Bayes*, *RBFNetwork_CFS*, *Logistic_CFS* e *Logistic*. O algoritmo *Logistic* obteve o melhor desempenho na medida de IS, o que indica que esse algoritmo possui a melhor taxa de precisão nesse corpus para classificar as sentenças corretas para compor o resumo.

No corpus do DUC 2001, o algoritmo *Logistic* apresentou o melhor resultado no R-1. Sua performance é estatisticamente similar aos algoritmos *AdaBoostM1*, *AdaBoostM1_CFS*, *Logistic_CFS*, *Multilayer Perceptron_CFS*, *SMO* e *SMO_CFS*. Com base no R-2, o algoritmo *SMO* obteve o melhor desempenho, mas seus resultados são estatisticamente similares com os algoritmos *AdaBoostM1*, *AdaBoostM1_CFS*, *Logistic*, *Logistic_CFS*, *Multilayer Perceptron_CFS* e *SMO_CFS*.

No corpus DUC 2002, o algoritmo *Logistic* obteve o melhor desempenho nas medidas do R-1 e R-2. Em termos do R-1, sua performance foi estatisticamente superior aos demais algoritmos, com exceção do *AdaBoostM1_CFS*. Já em relação ao R-2, o algoritmo *Logistic* apresentou resultados estatisticamente similares com os algoritmos *AdaBoostM1*, *AdaBoostM1_CFS*, *Logistic_CFS*, *Multilayer Perceptron*, *SMO* e *SMO_CFS*.

Sumarização Multidocumento

Na Tabela 7 são apresentados os resultados dos algoritmos de AM para classificação das sentenças na sumarização multidocumento. No corpus do DUC 2001, o *SMO* obteve a melhor performance com base nas duas medidas do R-1 e R-2. Seu desempenho com base no R-1 pode ser considerado estatisticamente similar aos algoritmos *SMO_CFS*, *AdaBoostM1*, *AdaBoostM1_CFS* e *RandomTree*. Levando em consideração o R-2, o algoritmo SMO não apresentou diferença estatística comparado com o *AdaBoostM1*, *AdaBoostM1_CFS*, *SMO_CFS* e *RandomTree*.

No DUC 2002, o algoritmo *AdaBoostM1* e sua variação *AdaBoostM1_CFS* obtiveram os melhores resultados no R-1 e R-2, respectivamente. O algoritmo *AdaBoostM1* não apresentou melhorias significativas em termos da medida do R-1 em relação aos algoritmos *MultilayerPerceptron_CFS*, *Logistic*, *Logistic_CFS*, *RandomTree_CFS*, *AdaBoostM1_CFS*, *J48*, *SMO* e *SMO_CFS*. Com base na medida do R-2, o algoritmo AdaBo-

Tabela 6 – Resultados (%) e desvio padrão entre parênteses dos algoritmos de AM para classificação das sentenças na sumarização monodocumento. O algoritmo com melhor performance em cada corpus é destacado em negrito e o grupo de algoritmos estatisticamente similar, se existir, é indicado usando †.

Classificadores	CNN			DUC 01		DUC 02	
	IS	R-1	R-2	R-1	R-2	R-1	R-2
AdaBoostM1	31,14	53,29 (20,75)	39,21 (24,40)	44,33† (9,65)	19,54† (10,98)	47,24 (8,93)	22,39† (9,63)
AdaBoostM1_CFS	31,18	53,70 (20,59)	39,50 (24,45)	44,38† (9,96)	19,66† (11,28)	47,37† (9,06)	22,52† (9,69)
IBK	20,60	44,22 (19,01)	27,14 (22,13)	43,18 (10,00)	18,10 (11,25)	46,22 (9,25)	21,01 (9,96)
IBK_CFS	21,29	44,75 (19,01)	28,06 (22,39)	42,76 (10,11)	17,63 (11,28)	45,56 (9,08)	20,55 (9,73)
J48	17,15	38,93 (19,33)	23,89 (21,43)	37,73 (11,77)	15,26 (10,24)	40,69 (11,79)	18,05 (9,74)
J48_CFS	17,92	40,11 (19,35)	24,76 (21,50)	37,70 (11,05)	15,09 (9,35)	41,77 (10,35)	18,47 (9,36)
Logistic	**32,54**	55,83† (20,78)	41,58† (24,65)	**44,65** (9,78)	19,82† (11,16)	**47,76** (8,93)	**22,67** (9,72
Logistic_CFS	32,17	56,04† (20,62)	**41,67** (24,60)	44,56† (9,79)	19,80† (11,32)	47,38† (8,97)	22,40† (9,61)
Multilayer Perceptron	31,67	54,27 (20,76)	39,91 (24,75)	43,81 (8,97)	18,93 (10,32)	47,04 (9,86)	22,19† (10,47)
Multilayer Perceptron_CFS	32,36	54,94 (20,91)	40,95† (24,87)	44,40† (9,86)	19,82† (11,33)	46,77 (9,28)	21,97 (9,73)
Naive Bayes	28,15	56,17† (20,24)	39,30 (25,30)	42,77 (10,13)	17,96 (11,08)	45,59 (9,57)	20,60 (9,77)
Naive Bayes_CFS	28,13	55,36 (20,24)	38,74 (25,00)	43,18 (9,77)	18,34 (10,98)	45,92 (9,70)	20,98 (9,81)
Random Forest	12,36	32,88 (18,60)	25,74 (21,69)	31,20 (13,41)	14,15 (9,80)	35,75 (13,43)	17,01 (8,92)
Random Forest_CFS	13,68	32,11 (18,78)	24,93 (21,66)	31,94 (13,52)	14,18 (10,50)	36,82 (13,09)	17,37 (9,39)
Random Tree	17,93	38,44 (19,50)	25,19 (21,89)	36,32 (12,77)	15,06 (10,09)	39,31 (12,11)	17,56 (9,62)
Random Tree_CFS	17,66	38,27 (19,41)	24,77 (21,60)	36,45 (11,84)	14,99 (10,13)	39,29 (12,52)	17,72 (9,81)
RBF Network	28,14	**56,22** (20,25)	39,44 (25,41)	42,39 (10,31)	17,38 (11,19)	45,46 (9,25)	20,52 (9,87)
RBF Network_CFS	28,70	55,96† (20,18)	39,51 (25,19)	43,45 (9,74)	18,64 (10,86)	46,46 (8,99)	21,33 (9,51)
SMO	28,58	48,65 (21,68)	35,40 (25,03)	44,48† (9,73)	**19,97** (11,33)	47,22 (8,99)	22,35† (9,80)
SMO_CFS	28,80	48,86 (21,66)	35,66 (24,95)	44,46† (9,79)	19,96† (11,42)	47,29 (8,99)	22,42† (9,82)

Tabela 7 – Resultados (%) e desvio padrão entre parênteses dos algoritmos de AM para classificação das sentenças na sumarização multidocumento. O algoritmo com melhor performance em cada corpus é destacado em negrito e o grupo de algoritmos estatisticamente similar, se existir, é indicado usando †.

Classificadores	DUC 01		DUC 02		DUC 03		DUC 04	
	R-1	R-2	R-1	R-2	R-1	R-2	R-1	R-2
AdaBoostM1	31,70† (6,61)	6,58† (5,47)	**34,17** (5,29)	7,58† (3,92)	35,06 (4,70)	7,62 (3,24)	35,78 (4,69)	8,28 (2,61)
AdaBoostM1_CFS	31,97† (6,70)	6,58† (5,93)	33,77† (5,32)	**7,62** (4,31)	35,06 (4,70)	7,72 (3,26)	36,60 (4,94)	8,89 (2,79)
IBk	27,42 (5,60)	3,71 (2,74)	31,18 (5,66)	5,33 (2,66)	35,01 (6,70)	6,32 (3,56)	35,19 (4,41)	7,25 (2,92)
IBk_CFS	28,94 (5,92)	4,32 (2,36)	31,38 (5,47)	5,76 (3,06)	34,86 (7,05)	7,03 (4,03)	34,67 (5,14)	7,49 (3,21)
J48	29,00 (7,11)	4,76 (4,07)	33,77† (6,11)	7,11† (4,78)	34,31 (6,90)	6,78 (4,20)	35,19 (5,14)	7,26 (3,05)
J48_CFS	28,91 (5,74)	3,99 (2,72)	31,02 (4,77)	5,17 (2,59)	31,91 (8,74)	5,86 (3,42)	33,07 (4,34)	6,36 (2,64)
Logistic	28,16 (6,97)	5,02 (4,18)	34,14† (6,39)	7,60† (4,03)	36,56† (6,28)	8,03† (4,58)	37,09†± 4,80)	9,02† (3,06)
Logistic_CFS	27,31 (7,61)	4,63 (3,92)	34,08† (6,34)	7,57† (4,00)	35,35 (7,90)	8,19† (4,83)	**37,48** (5,33)	**9,46** (3,21)
MultilayerPerceptron	28,21 (5,57)	4,29 (2,71)	32,05 (6,30)	6,46 (3,27)	35,67 (7,50)	8,12† (4,75)	36,33 (4,98)	8,34 (2,79
MultilayerPerceptron_CFS	28,39 (8,10)	4,42 (3,40)	34,16† (5,90)	7,21 (3,81)	35,27 (6,93)	7,76 (4,95)	36,89 (4,57)	8,66 (2,85)
NaiveBayes	26,67 (7,40)	4,09 (3,78)	31,14 (8,14)	6,02 (3,86)	36,29† (7,60)	**8,45** (5,31)	37,29†4,32)	9,23† (3,04)
NaiveBayes_CFS	27,58 (7,64)	4,63 (4,08)	31,61 (7,92)	6,43 (3,67)	36,23† (7,23)	7,55† (5,25)	37,11† (4,64)	9,24† (3,17)
RandomForest	25,04 (6,31)	5,00 (3,60)	32,99 (6,14)	6,58 (4,09)	18,34 (8,01)	3,73 (2,32)	34,03 (8,10)	8,39 (3,32)
RandomForest_CFS	25,88 (6,83)	4,24 (3,27)	33,46 (5,87)	7,07 (3,14)	22,42 (7,24)	4,69 (1,96)	36,18 (4,63)	8,87 (2,95)
RandomTree	30,74† (7,84)	5,96† (3,88)	33,83 (4,94)	6,79 (3,27)	31,92 (7,80)	6,56 (3,84)	36,80 (4,73)	8,84 (3,04)
RandomTree_CFS	27,69 (5,27)	4,58 (2,76)	34,0† (5,25)	7,60† (3,97)	32,68 (8,07)	6,87 (4,61)	36,26 (5,30)	8,35 (3,43)
RBFNetwork	26,29 (7,03)	3,91 (3,49)	31,19 (8,27)	5,79 (3,75)	36,39† (7,34)	8,18† (5,56)	36,89 (4,52)	8,96 (3,22)
RBFNetwork_CFS	27,70 (7,85)	4,64 (4,08)	31,84 (7,69)	6,48 (3,67)	**37,13** (6,81)	7,93 (5,21)	37,24† (4,69)	9,28† (3,28)
SMO	**32,74** (6,31)	**7,13** (6,02)	33,66† (4,90)	7,05 (3,32)	33,97 (5,12)	7,15 (3,05)	35,40 (4,91)	8,03 (2,59)
SMO_CFS	32,08† (6,59)	6,45† (5,30)	33,66† (4,90)	7,05 (3,32)	34,98 (5,84)	7,64 (4,13)	35,43 (4,89)	8,01 (2,56)

ostM1_CFS apresentou resultados estatisticamente similares em relação aos algoritmos *Logistic*, *RandomTree_CFS*, *AdaBoostM1*, *Logistic_CFS* e *J48*.

No DUC 2003, os algoritmos *RBFNetwork_CFS* e *Naive Bayes* obtiveram o melhor desempenho nas medidas do R-1 e R-2, respectivamente. Com base no R-1, o algoritmo RBFNetwork_CFS apresentou resultados estatisticamente similar em relação aos algoritmos *Logistic*, *RBFNetwork*, *NaiveBayes* e *NaiveBayes_CFS*. Já em relação ao R-2, o algoritmo Naive Bayes não obteve uma diferença significativa em relação aos algoritmos *Logistic_CFS*, *RBFNetwork*, *Multilayer Perceptron* e *Logistic*.

Por fim, no DUC 2004, o algoritmo *Logistic_CFS* obteve o melhor desempenho nas duas medidas do R-1 e R-2. Contudo, em ambos os casos, ele não apresentou melhorias significativas em relação aos algoritmos *Logistic*, *NaiveBayes*, *NaiveBayes_CFS* e *RBFNetwork_CFS*.

Com base nos resultados dos experimentos realizados em ambas as tarefas monodocumento e multidocumento, algumas conclusões podem ser delineadas:

- Como esperado, o algoritmo CFS reduziu a dimensionalidade das características utilizadas, mas isso, em geral, ligeiramente influenciou nas medidas do R1 e R2 em ambas as tarefas. Em 57,14% das 70 comparações realizadas nos experimentos desta seção, o desempenho dos algoritmos utilizando somente as características selecionadas pelo algoritmo CFS obtiveram melhores resultados do que usar todas as características. Nos casos em que o algoritmo CFS levou a uma diminuição na performance, em 95,71% deles não houve diferença estatística ao nível de 95% de nível de confiança.
- Na sumarização monodocumento, os algoritmos *Logistic_CFS* e *RBFNetwork_CFS* apresentaram o melhor desempenho global em termos da medida do R-1. Os métodos de pontuação selecionados pelo algoritmo CFS em todos os corpora foram: *bushy path*, entidades nomeadas, dados numéricos, centralidade sobreposição, centralidade do cosseno, posição das sentenças versão 2, similaridade com o título, *textrank*, FT-FIS e frequência das palavras.
- Na sumarização multidocumento, os algoritmos *AdaBoostM1_CFS* e *Logistic* obtiveram o melhor resultado global em todos os corpora com base no R-1. Os métodos de pontuação selecionados pelo algoritmo CFS em todos os corpora foram: similaridade agregada, similaridade léxica, centralidade sobreposição, centralidade cosseno, posição das sentenças versão 2, similaridade com o título, *textrank*, FT-FIS, coocorrência e frequência das palavras.
- Analisando os resultados, é possível observar que os algoritmos *Logistic* e *AdaBoostM1* alcançaram uma alta taxa de cobertura na identificação de sentenças que deveriam ser incluídas nos resumos. Em geral, a maioria dos algoritmos adotados classificam mais sentenças do que as permitidas pela taxa de compressão como sendo

da classe Resumo. Contudo, a estratégia de usar a confiança dos classificadores para discriminar as sentenças classificadas como sendo da classe Resumo, levando em consideração a taxa de compressão, não foi efetiva. Na maioria dos casos isso acabou levando a escolha de sentenças não tão adequadas. Tal problema acabou diminuindo a performance do processo de sumarização em relação as medidas de avaliação do R-1, R-2 e IS.

- Uma grande sobreposição entre os exemplos da classe Resumo e Não-Resumo foi observada, isto é, sentenças de classes diferentes, mas com características muito similares. Esse comportamento dificultou ainda mais o processo de classificação.

3.2.6 Comparando os Resultados com Sistemas do Estado da Arte

Nesta seção, são realizadas comparações entre o melhor método individual identificado na Subseção 3.2.3, a combinação que obteve melhor performance na Subseção 3.2.4, e o melhor algoritmo de AM identificado na Subseção 3.2.5, com diversos sistemas do estado da arte nas tarefas de sumarização monodocumento e multidocumento.

Sumarização Monodocumento

Na tarefa de sumarização monodocumento, os resultados obtidos foram comparados com: **(i)** Os melhores sistemas participantes das competições do DUC 2001 e 2002 identificados na execução do ROUGE neste trabalho. Esses sistemas são rotulados originalmente como T (DUC 2001) e 28 (DUC 2002); e **(ii)** Três sistemas de sumarização que apresentaram melhor performance na avaliação comparativa descrita em Batista et al. (2015). Esses sistemas são descritos a seguir:

- **Autosummarizer** (AUTOSUMMARIZER, 2016) é um sumarizador monodocumento disponível online, que seleciona as sentenças mais importantes do documento de entrada para gerar um resumo. Infelizmente, detalhes de como esse sistema funciona não foram encontrados. Contudo, ele apresentou bons resultados em uma análise comparativa entre diferentes sistemas de sumarização (BATISTA et al., 2015).

- **Classifier4J** (LOTHIAN, 2003) é uma biblioteca que fornece serviços para a classificação e sumarização monodocumento de textos. Classifier4J primeiro extrai as cem palavras mais frequentes do documento de entrada como palavras-chave, e então seleciona as primeiras sentenças do texto que possuem no mínimo uma dessas palavras-chave extraídas.

- **HP-UFPE Functional Summarization** (HP-UFPE FS) (FERREIRA et al., 2014) é um sistema de sumarização baseado na combinação de métodos de pontuação

superficiais. Para mensurar a importância das sentenças esse sistema utiliza a combinação das seguintes técnicas: FT-FIS, Similaridade Léxica, Posição das Sentenças versão 1 e Similaridade com o Título.

Na Tabela 8 são apresentados os resultados da comparação entre os melhores métodos e combinações identificados neste trabalho com base na medida do R-1, em relação aos trabalhos relacionados citados anteriormente.

Tabela 8 – Resultados (%) e desvio padrão entre parênteses dos sistemas na sumarização monodocumento. O sistema com melhor performance em cada corpus é destacado em negrito e o grupo de algoritmos estatisticamente similar, se existir, é indicado usando o símbolo †.

Sistemas	**CNN**		
	IS	R-1	R-2
AutoSummarizer	23,16	48,81 (18,70)	32,74 (22,70)
Classifier4J	23,89	46,63 (20,32)	32,15 (23,13)
Combinação Linear A	**29,89**	**57,93** (20,22)	**41,54** (25,32)
FT-FIS	23,99	54,25 (19,87)	35,65 (25,44)
HP-UFPE FS	24,75	50,71 (20,34)	34,58 (24,38)
RBFNetwork	28,14	56,22 (20,25)	39,44 (25,41)
Sistemas	**DUC 2001**		
	IS	R-1	R-2
AutoSummarizer	-	41,92 (9,04)	16,63 (9,95)
Classifier4J	-	44,44† (9,85)	19,86† (11,34)
Comb. Linear Ponderada A	-	**44,98** (9,56)	20,12† (11,04)
HP-UFPE FS	-	35,91 (11,78)	11,78 (9,78)
Logistic	-	44,65† (9,78)	19,82† (11,16) -
Posição das Sentenças versão 2	-	43,75 (10,47)	19,57† (11,64)
Sistema T	-	44,53† (9,23)	**20,27** (10,75)
Sistemas	**DUC 2002**		
	IS	R-1	R-2
AutoSummarizer	-	43,79 (8,78)	19,17 (9,31)
Comb. Linear Ponderada A	-	47,73† (8,47)	*22,34*† (9,56)
Classifier4J	-	47,09 (8,93)	22,12 (9,87)
Logistic	-	47,76† (8,93)	22,67† (9,72)
HP-UFPE FS	-	45,70 (9,31)	20,59 (9,88)
Posição das Sentenças versão 2	-	46,94 (9,20)	22,14 (10,01)
Sistema 28	-	**48,07** (8,90)	**22,88** (9,96)

No corpus CNN, a melhor combinação identificada (combinação linear A) obteve a melhor performance em termos do R-1, R-2, e da medida IS. Os resultados obtidos pela combinação linear A apresentaram uma melhoria significativa em relação aos demais sistemas ao nível de 95% de confiança. O melhor algoritmo de AM (RBFNetwork) e o melhor método individual (FT-FIS) também obtiveram resultados estatisticamente superiores do que os sistemas *AutoSummarizer*, *Classifier4J* e HP-UFPE FS. Com relação à medida IS, os resultados demonstram que ainda existe muito espaço para melhorias. O melhor resultado encontrado, obtido pela combinação linear A, foi de apenas 29,89%. Mesmo levando em consideração o melhor resultado obtido em todos os experimentos realizados, que é de 32,54% obtido pelo algoritmo *Logistic* na Subseção 3.2.5, ainda assim é muito

baixo. É importante salientar que todas as decisões relativas as escolhas dos métodos e combinações foram baseadas na medida de cobertura do R-1. Então, é possível que outras combinações não levadas em consideração neste trabalho possam levar a resultados melhores com base na medida IS.

Em relação ao corpus do DUC 2001, a combinação linear ponderada A obteve a melhor performance com base no R-1. Contudo, tal resultado não é estatisticamente superior ao sistema *Classifier4J*, ao melhor algoritmo de AM (*Logistic*) e ao melhor participante da competição do DUC 2001 (Sistema T). Em termos do R-2, o Sistema T apresentou a melhor performance, mas esse desempenho só foi significativamente superior que os sistemas *AutoSummarizer* e HP-UFPE FS.

No corpus do DUC 2002, o sistema participante do DUC 2002 (Sistema 28), o melhor algoritmo de AM *Logistic* e a combinação linear ponderada A apresentam os três melhores resultados com base no R-1 e R-2. Nenhuma diferença estatística entre eles foi observada ao nível de 95% de confiança. A combinação linear ponderada A apresentou melhorias significativas em relação ao melhor método individual (posição das sentenças versão 2) e aos sistemas *AutoSummarizer*, *Classifier4J* e HP-UFPE FS.

Surpreendentemente, mesmo depois de mais de uma década desde as competições do DUC 2001 e DUC 2002, os melhores participantes identificados nos experimentos realizados neste trabalho (Sistema T e Sistema 28) apresentaram resultados competitivos em relação a outros sistemas propostos mais recentemente. Essa boa performance se deve principalmente porque esses sistemas foram desenvolvidos especificamente para os corpora de suas respectivas competições.

Utilizando as melhores configurações identificadas neste capítulo não foi possível obter melhorias significativas em relação ao Sistema T e Sistema 28 nos corpora do DUC 2001 e 2002, respectivamente. Contudo, em ambos os corpora, as combinações lineares ponderadas identificas na Subseção 3.2.4 apresentaram resultados estatisticamente superiores do que o método de posição das sentenças versão 2, que foi usado como *baseline* nas competições originais do DUC, e também em relação aos sistemas *AutoSummarizer*, *Classifier4J* e HP-UFPE FS.

Na Tabela 9 são apresentados dois exemplos de resumos produzidos pelo sistema 28 e pela combinação linear ponderara A para o documento *AP880622-0184* do corpus do DUC 2002. Além disso, os dois resumos de referência criados pelos organizadores da competição do DUC 2002 também são apresentados. Os dois sistemas geraram resumos semelhantes e obtiveram um bom desempenho na medida do R-1: combinação linear ponderada A (70,79%) e o sistema 28 (61,38%).

Sumarização Multidocumento

Os resultados obtidos na sumarização multidocumento são comparadas com: **(i)** Os melhores sistemas participantes das competições do DUC 2001-2004 identificados nos

Tabela 9 – Resumos gerados pelo Sistema 28 e pela combinação linear ponderada A para o documento AP880622-0184 do corpus do DUC 2002. Os dois resumos de referência disponíveis também são listados.

Sistema/Combinação	Resumo
Sistema 28	Beverly Sills, Lauren Bacall, Betty Comden and Phyllis Newman are among performers who will sing, act and make guest appearances at a birthday bash in August for conductor Leonard Bernstein. The Leonard Bernstein Gala Birthday Performance is a benefit concert scheduled for the composer's 70th birthday, Aug. 25, to raise money for the Tanglewood Music Center, where Bernstein got his conducting start. The concert will celebrate Bernstein's accomplishments in popular music with excerpts from "West Side Story", "On the Town" and others. Dame Gwyneth Jones and Frederica von Stade will be among those performing highlights from "Fidelio", "Tristan und Isolde" and other works to honor Bernstein's landmark opera recordings.
Combinação Linear Ponderada A	Beverly Sills, Lauren Bacall, Betty Comden and Phyllis Newman are among performers who will sing, act and make guest appearances at a birthday bash in August for conductor Leonard Bernstein. The Leonard Bernstein Gala Birthday Performance is a benefit concert scheduled for the composer's 70th birthday, Aug. 25, to raise money for the Tanglewood Music Center, where Bernstein got his conducting start. Performances will include the Boston Symphony Orchestra, the Boston Pops Orchestra and the Tanglewood Festival Chorus under the direction of some of the many conductors whose careers have been guided by Bernstein. Bacall and soprano Barbara Hendricks will perform a movement from Bernstein's Symphony No. 3, "Kaddish".
Resumo de referência 1	The Leonard Bernstein Gala Birthday Performance is a benefit concert scheduled for the composer's 70th birthday, August 25, to raise money for the Tanglewood Music Center, where Bernstein started his conducting career. Beverly Sills will host. Lauren Bacall, Betty Comden, and Phyllis Newman are among the performers who will sing, act, and make guest appearances. The Boston Symphony Orchestra, the Boston Pops Orchestra, and the Tanglewood Festival Chorus, directed by some of the conductors mentored by Bernstein, will also perform. The concert will celebrate Bernstein's contributions to classical and popular compositions and landmark opera recordings. Tickets cost 20*to*5000.
Resumo de referência 2	Beverly Sills, Lauren Bacall, Betty Comden, and Phyllis Newman will appear at the Leonard Bernstein Gala Birthday Performance in August. The concerts mark his 70th birthday and will benefit the Tanglewood Music Center, where he got his start as a conductor. Conductors who were mentored by Bernstein will direct the Boston Symphony, the Boston Pops Orchestra, and the Tanglewood Festival Chorus. Bernstein's compositions will be honored by selections from "Kaddish", "Serenade", "On the Town", and "West Side Story." Bernstein's landmark opera recordings will also be honored. Tickets go from $20 on the lawn to $5,000 for benefactors.

experimentos realizados neste capítulo com base na medida do ROUGE-1; e **(ii)** Com os seguintes sistemas do estado da arte ICSISumm (GILLICK et al., 2009), Greedy-KL (HAGHIGHI; VANDERWENDE, 2009), LLRSum (CONROY; SCHLESINGER; O'LEARY, 2006), ProbSum (NENKOVA; VANDERWENDE; MCKEOWN, 2006) e Sume (BOUDIN; MOUGARD; FAVRE, 2015). Os resumos do sistema Sume foram gerados utilizando a própria implementação do sistema disponibilizada pelos autores em [5]. Os resumos dos outros sistemas foram gerados e disponibilizados por Hong, Marcus e Nenkova (2015).

Na Tabela 10 são apresentados os resultados comparativos entre os melhores métodos, combinações e algoritmos de AM identificados neste trabalho com base na medida R-1, em relação aos trabalhos relacionados citados anteriormente.

Tabela 10 – Resultados (%) e desvio padrão entre parênteses dos sistemas na sumarização multidocumento. O sistema com melhor performance em cada corpus é destacado em negrito e o grupo de algoritmos estatisticamente similar, se existir, é indicado usando o símbolo †.

Sistemas	**DUC 2001**		**DUC 2002**	
	R-1	R-2	R-1	R-2
Melhor Combinação	33,63† (7,79)	7,67† (6,10)	35,83 (5,06)	7,83 (3,59)
Melhor Participante DUC	31,69 (6,43)	6,30† (3,76)	35,21 (5,30)	7,66 (3,30)
Melhor Método Individual	30,63 (6,66)	5,94 (3,74)	33,78 (5,65)	7,16 (3,69)
Melhor Algoritmo AM	32,74† (6,31)	7,13† (6,02)	34,17 (5,29)	7,58 (3,92)
Greedy-KL	32,84† (6,43)	6,70† (3,64)	35,79 (5,74)	7,49 (3,61)
ICSISumm	**33,88** (6,95)	**7,75** (4,30)	**37,34** (5,05)	**9,53** (3,83)
LLRSum	32,00† (5,88)	6,76† (3,25)	32,84 (5,55)	6,75 (3,72)
ProbSum	29,73 (5,41)	5,16 (2,64)	32,57 (4,74)	7,06 (3,63)
Sume	33,37† (7,14)	7,73 (4,29)	34,30 (5,26)	8,15 (3,92)
Sistemas	**DUC 2003**		**DUC 2004**	
	R-1	R-2	R-1	R-2
Melhor Combinação	39,61† (9,14)	9,76† (5,39)	**38,58** (4,23)	**9,80** (2,98)
Participante DUC	38,53† (7,98)	9,17† (5,59)	37,69† (4,08)	8,98† (3,08)
Melhor Método Individual	37,61 (8,00)	9,42† (5,43)	36,85 (4,97)	8,34 (3,75)
Melhor Algoritmo AM	37,13 (6,81)	7,93 (5,21)	37,48 (5,33)	9,46† (3,21)
Greedy-KL	39,92† (7,82)	8,82 (4,80)	38,27† (4,73)	8,96 (3,09)
ICSISumm	**40,03** (8,05)	**11,06** (6,15)	38,42† (4,14)	**9,80** (3,17)
LLRSum	36,68 (7,74)	7,99 (4,27)	35,90 (5,01)	8,06 (3,12)
ProbSum	36,09 (8,27)	8,94 (3,73)	35,37 (4,41)	8,18 (3,00
Sume	39,34† (7,34)	9,73† (5,51)	37,29 (4,24)	8,83 (2,71)

No corpus do DUC 2001, os dois melhores resultados com base no R-1 foram obtidos pelo sistema ICSISumm e pela combinação linear A. Eles obtiveram resultados estatisticamente superiores em relação aos sistemas Greedy-KL, LLRSum, Sume e ao algoritmo SMO. Somente o sistema ICSISumm apresentou uma melhora significativa em relação ao Sistema P (melhor participando do DUC 2001). A combinação linear A apresentou

[5] https://github.com/boudinfl/sume

melhorias significativas em relação ao sistema ProbSum e ao método de similaridade com o título. Em termos da medida R-2, ICSISumm obteve a melhor performance, contudo, ele só apresentou uma melhoria significativa em relação ao sistema ProbSum e ao método de similaridade com o título.

Nos experimentos realizados no corpus do DUC 2002, o sistema ICSISumm obteve a melhor performance em relação às medidas do R-1 e R-2, apresentando melhorias significativas em comparação com todos os demais sistemas. Com base no R-1, a combinação linear A obteve resultados estatisticamente superiores em relação aos sistemas LLRSum, ProbSum, Sume, a técnica de posição das sentenças versão 2 e ao algoritmo AdaBoostM1_CFS. Apenas o sistema ICSISumm obteve um desempenho significativamente melhor do que o Sistema 26 (melhor participando do DUC 2002).

No corpus do DUC 2003, o sistema ICSISumm mais uma vez obteve o melhor desempenho em relação às medidas do R-1 e R-2. Contudo, baseado no R-1, seu desempenho é estatisticamente similar em relação à combinação linear ponderada B, aos sistemas Greedy-KL e Sume, e ao Sistema 12 (melhor participando do DUC 2003). A combinação linear ponderada B apresentou uma performance significativamente superior com base no R-1 em comparação com os sistemas LLRSum e ProbSum, o algoritmo RBFNetwork_CFS (melhor algoritmo de AM identificado para esse corpus) e o método de centralidade do cosseno. Nenhum dos sistemas avaliados obteve resultados estatisticamente superiores em relação ao Sistema 12. Em termos do R-2, o sistema ICSISumm apresenta um desempenho estatisticamente semelhante com a combinação linear ponderada B, o método de centralidade do cosseno e o Sistema 12.

No corpus do DUC 2004, a melhor performance com base no R-1 e R-2 foi obtida pela combinação linear A. O sistema ICSISumm também obteve a melhor performance em termos do R-2. Com base no R-1, o desempenho da combinação linear A foi estatisticamente superior aos sistemas LLRSum, ProbSum, Sume, ao método *bushy path*, e ao algoritmo Logistic_CFS. Assim como no corpus do DUC 2003, nenhum sistema apresentou melhorias significativas em termos do R-1 em relação ao sistema CLASSY 04 (Sistema 65) (melhor sistema participante do DUC 2004) (CONROY et al., 2004). Com base no R-2, a combinação linear A e o sistema ICSISumm obtiveram desempenho estatisticamente melhores do que o método *bushy path* e os sistemas Greed-KL, LLRSum, ProbSum e Sume.

O sistema ICSISumm claramente se destaca dos demais sistemas do estado da arte comparados, apresentando os melhores resultados em praticamente todos os corpora. Apesar dos avanços mais recentes na sumarização multidocumento, o desempenho dos participantes originais das competições do DUC ainda é muito competitivo. Somente o sistema ICSISumm conseguiu obter um desempenho estatisticamente superior ao nível de 95% de confiança nos corpora do DUC 2001 e 2002 com base no R-1, e somente no DUC 2002 em termos do R-2.

As estratégias de combinação investigadas, na maioria dos casos, apresentaram melhorias significativas em relação aos melhores métodos de pontuação de sentenças utilizados individualmente. Tanto na sumarização monodocumento quanto na multidocumento, foi possível obter um desempenho competitivo com sistemas do estado da arte analisados através da combinação de técnicas de pontuação de sentenças superficiais e utilizando uma abordagem simples para evitar redundância.

3.3 Considerações Finais do Capítulo

Neste capítulo, foi apresentada uma extensa análise comparativa entre diversas técnicas de pontuação de sentenças superficiais para a sumarização monodocumento e multidocumento no domínio de artigo de notícias. Esses métodos de pontuação foram avaliados em vários cenários de aplicação: individualmente, usando estratégias de combinação e comparando os melhores resultados com diversos sistemas do estado da arte, utilizando as medidas tradicionais de avaliação.

Os resultados obtidos demonstram que o desempenho individual dos métodos é razoável, e que a estratégia de combiná-los leva à obtenção de uma melhora significativa na maioria dos cenários identificados. Além disso, os experimentos demonstram que é possível identificar combinações que apresentam desempenho competitivo comparado com diversos sistemas do estado da arte. As combinações com melhor desempenho exploram a diversidade e o desempenho dos melhores métodos individuais para otimizar a tarefa de pontuação de sentenças. No entanto, nenhum padrão em tais combinações foi encontrado de forma a indicar que uma mesma combinação obtém os melhores resultados em todos os corpora avaliados.

Com base na medida do ROUGE-1, e levando em consideração todos os experimentos realizados, podemos apontar que os dez melhores métodos de pontuação de sentenças superficiais são: similaridade agregada, *bushy path*, entidades nomeadas, sintagmas nominais e verbais, centralidade do cosseno, frequência das palavras, posição das sentenças versão 2, similaridade com o título, *textrank* e FT-FIS. Em relação às estratégias de combinação, a combinação linear e a combinação linear ponderada apresentam melhor desempenho do que as combinações baseadas em voto majoritário, *condorcet* e utilizando algoritmos de aprendizagem de máquina.

Os resultados obtidos neste capítulo permitem: **(i)** Identificar quais aspectos influenciam em cada uma das tarefas de sumarização (monodocumento e multidocumento); **(ii)** Analisando os resultados gerados pelas técnicas, combinações e sistemas investigados neste capítulo, é possível observar um alto desvio padrão nos resultados das medidas do ROUGE-1 e ROUGE-2. Isso demonstra que nenhum dos métodos, combinações e sistemas analisados consegue manter um alto desempenho para todos os documentos ou grupos de documentos.

Os sistemas ICSISumm e Sume, que são baseados na maximização da cobertura de conceitos relevantes usando PLI, apresentaram bons resultados nos experimentos realizados neste capítulo. Revisando a literatura, observou-se que este tipo de abordagem tem apresentado resultados competitivos com o estado da arte para a sumarização multidocumento. Contudo, no melhor do conhecimento do autor desta tese, nenhum trabalho que explora esse tipo de abordagem foi identificado para sumarização monodocumento. Com base nessa lacuna, no próximo capítulo, apresentaremos a proposta de uma abordagem baseada em conceitos para sumarização monodocumento utilizando PLI que leva em consideração os aspectos de informatividade, redundância e coesão local dos resumos gerados.

4 UMA ABORDAGEM BASEADA EM CONCEITOS UTILIZANDO PLI PARA A SUMARIZAÇÃO MONODOCUMENTO

O objetivo de um sistema de SAT monodocumento é identificar e extrair as informações mais relevantes a partir de um documento de entrada e apresentá-las de forma resumida (NENKOVA; MCKEOWN, 2012). Neste sentido, um sistema de SAT completo deve considerar os seguintes aspectos: (i) **Informatividade** - O resumo gerado deve conter as informações mais relevantes do documento original; (ii) **Redundância** - Sobreposição de informações entre as sentenças do resumo deve ser evitada; e (iii) **Coesão** - Já que o resumo produzido, em geral, é destinado para a leitura humana, ele deve ser coeso e gramaticalmente correto.

As abordagens extrativas selecionam o subconjunto de sentenças originais mais relevantes do documento de entrada e as utilizam para compor o resumo final. Supondo que o documento de entrada está gramaticalmente correto, os resumos gerados também estarão corretos, já que eles são gerados utilizando as sentenças originais do documento. Contudo, resumos produzidos por abordagens extrativas, em geral, apresentam problemas de coesão, por exemplo, correferências em aberto e quebra no fluxo de ideias entre as sentenças. Atualmente, a maioria das abordagens propostas para sumarização monodocumento tratam de aspectos relacionados a informatividade e a redundância (MIHALCEA; TARAU, 2004; FERREIRA et al., 2013; GARCÍA-HERNÁNDEZ; LEDENEVA, 2013), enquanto que poucos trabalhos abordam a coesão dos resumos gerados (PARVEEN; RAMSL; STRUBE, 2015a; PARVEEN; STRUBE, 2015b).

Abordagens baseadas na ideia de maximizar a cobertura de conceitos importantes utilizando Programação Linear Inteira (PLI) (GILLICK et al., 2009; BOUDIN; MOUGARD; FAVRE, 2015; CAO et al., 2015; LI; LIU; ZHAO, 2015) vêm sendo muito investigadas recentemente, especialmente para sumarização multidocumento. Em tais abordagens, o processo de sumarização é tratado como um *problema de maximização de cobertura*, ou seja, selecionar um subconjunto de sentenças do documento original que maximize uma função objetivo sob certas restrições. Embora tal problema tenha sido demonstrado ser NP-difícil, soluções aproximadas ou exatas têm sido encontradas utilizando PLI (MCDONALD, 2007). Abordagens baseadas na maximização de conceitos são particularmente interessantes, pois, solucionam conjuntamente a informatividade, selecionando os conceitos mais relevantes do documento de entrada, e indiretamente evitam redundância, já que as sentenças só são selecionadas se possuírem conceitos novos e relevantes para o resumo gerado.

A análise apresentada no Capítulo 3 evidenciou o bom desempenho na tarefa de suma-

rização multidocumento obtido pelos sistemas baseados em conceitos ICSISumm e Sume. Alguns trabalhos que modelam a tarefa de sumarização como um problema de otimização combinatória foram identificados para a sumarização monodocumento (HIRAO et al., 2013; KIKUCHI et al., 2014; PARVEEN; RAMSL; STRUBE, 2015a; PARVEEN; STRUBE, 2015b; DURRETT; BERG-KIRKPATRICK; KLEIN, 2016). Esses trabalhos tratam a tarefa de seleção de frases como um problema de otimização, buscando maximizar diferentes aspectos como informatividade e coesão dos resumos gerados. Contudo, nenhum desses trabalhos utiliza uma abordagem baseada em conceitos seguindo a modelagem proposta por Gillick et al. (2009).

Baseado nesta lacuna, neste capítulo é apresentada a proposta de uma abordagem baseada em conceitos adotando PLI para a sumarização monodocumento que leva em consideração os aspectos de **Informatividade**, **Redundância** e **Coesão** do resumo gerado. O grau de informatividade dos resumos é estimado através da combinação de duas características: *Posição* e *Frequência* das sentenças. Para isso, frases recebem uma maior probabilidade de serem selecionadas caso estejam no início do documento e também possuam conceitos relevantes, ou caso estejam no meio ou no fim do documento e introduzam novos conceitos para o resumo gerado. Para evitar a inclusão de redundância nos resumos gerados, duas estratégias são usadas: **(i)** Indiretamente pelo uso do modelo de PLI baseado em conceitos que busca inserir a maior quantidade de conceitos diversos, respeitando o tamanho máximo do resumo a ser gerado; e **(ii)** Para garantir que sentenças redundantes não sejam incluídas nos resumos, utilizou-se a mesma estratégia adotada no Capítulo 3 de somente incluir uma nova sentença no resumo caso ela não possua uma similaridade maior que um dado limiar com nenhuma outra sentença já presente no resumo. Finalmente, para tratar a coesão do resumo gerado, duas estratégias são adotadas **(i)** A inclusão de restrições no modelo de PLI para evitar a ocorrência de correferências em aberto e quebras no fluxo de discurso entre as frases do resumo usando conectivos explícitos de discurso; e **(ii)** Integração do método de Grafo de Entidades (GUINAUDEAU; STRUBE, 2013) para estimar a coesão local do resumo gerado.

As principais contribuições deste capítulo são:

- Um novo método para a ponderação da relevância dos conceitos que combina os aspectos de posição e frequência das sentenças. Além disso, a importância dos conceitos adjacentes também é considerada durante o processo.

- Uma nova estratégia para a distribuição dos pesos dos conceitos que prioriza sentenças no início do documento caso elas possuam conceitos relevantes e sentenças subsequentes, se elas introduzirem novos conceitos para o resumo gerado.

- Investigação do impacto nas medidas de cobertura do ROUGE-1 e ROUGE-2 da inclusão de um conjunto de restrições ao modelo de PLI com o objetivo de evitar a presença de correferências em aberto e quebras no fluxo de discurso dos resumos

gerados. Além disso, o modelo de Grafo de Entidades proposto por Guinaudeau e Strube (2013) também foi avaliado. Apesar de ser um aspecto muito importante, a avaliação da coesão dos resumos demanda muito tempo, já que precisa ser realizada manualmente. Como o foco deste trabalho é na seleção de conteúdo relevante (informatividade), deixamos a avaliação da coesão dos resumos gerados como trabalho futuro.

O restante deste capítulo está organizado como segue: A Seção 4.1 descreve as etapas da abordagem proposta. Na Seção 4.2 são apresentados os resultados dos experimentos realizados para avaliar diferentes aspectos da abordagem proposta. Por fim, na Seção 4.3 são apresentadas as conclusões deste capítulo.

4.1 Abordagem Proposta

Em uma abordagem baseada na estratégia de maximização de conceitos importantes, a tarefa de seleção de sentenças é tratada como um problema de máxima cobertura, ou seja, selecionar o subconjunto de sentenças que maximize a cobertura de conceitos relevantes do documento de entrada, levando em consideração o tamanho máximo do resumo que deve ser gerado (GILLICK et al., 2009). Na abordagem proposta neste capítulo, além da informatividade, a coesão do resumo também é levada em consideração. Para isso, esses dois aspectos são tratados simultaneamente utilizando o modelo de otimização apresentado na Equação 4.1.

$$MAX \quad \sum_{c_i \in C} w_i c_i + \sum_{s_j \in S} co_j s_j \tag{4.1a}$$

$$s.t. \quad \sum_{s_j \in S} l_j s_j \leq L \tag{4.1b}$$

$$s_j Occ_{ij} \leq c_i \quad \forall i, j \tag{4.1c}$$

$$\sum_{s_j \in S} s_j Occ_{ij} \geq c_i \quad \forall i, j \tag{4.1d}$$

$$Ds_j \leq \sum_{s_d \in S_D} s_d \quad \forall j, d \tag{4.1e}$$

$$\sum_{s_r \in S_r} s_r \leq 0 \tag{4.1f}$$

$$c_i, s_j, s_d, s_r, Occ_{ij} \in \{0, 1\} \quad \forall i, j, d, r \tag{4.1g}$$

Nessa equação, as variáveis binárias c_i, s_j e Occ_{ij} indicam o conceito (c_i), a sentença (s_j), e a presença do conceito c_i na sentença s_j, respectivamente. A variável w_i é o peso (importância) de cada conceito c_i pertencente ao conjunto de conceitos C extraído do documento de entrada. A variável co_j é a pontuação da coesão local das sentenças gerada

adotando o modelo de Grafo de Entidades (GUINAUDEAU; STRUBE, 2013). A primeira parte da função objetivo $\sum_{c_i \in C} w_i c_i$ garante a informatividade do resumo, selecionando o maior número de conceitos importantes, enquanto que a segunda parte é relacionada a coesão local $\sum_{s_j \in S} co_j s_j$. A variável l_j é o tamanho de cada sentença s_j do conjunto de sentenças S contidas no documento e L é o tamanho máximo do resumo que deve ser gerado. As Equações 4.1c e 4.1d representam as restrições que devem ser atendidas para assegurar a consistência do modelo, ou seja, se uma frase é selecionada, isto implica em selecionar todos os conceitos que ela contém, e um conceito só é selecionado se ele está presente em, pelo menos, uma sentença selecionada. A Equação 4.1e é a restrição introduzida neste trabalho para evitar os problemas de correferências em aberto e quebras no fluxo de discurso entre as sentenças. Essa restrição garante que, se uma sentença s_j possui uma dependência com D outras sentenças, s_j só deve ser inserida no resumo se todas as D sentenças (S_D) de que ela depende também forem selecionadas para compor o resumo. Por fim, a Equação 4.1f garante que todas as sentenças S_r removidas durante a etapa de filtragem, explicada posteriormente, não podem ser selecionadas para compor o resumo gerado.

Duas questões fundamentais que devem ser definidas em uma abordagem baseada em conceitos são: **(i)** encontrar uma representação adequada para os conceitos; e **(ii)** definir um método para estimar a relevância de um conceito que justifique sua inclusão no resumo. Este trabalho adota bigramas como conceitos, devido aos resultados encorajadores relatados em trabalhos anteriores (GILLICK et al., 2009; BOUDIN; MOUGARD; FAVRE, 2015; LI; LIU; ZHAO, 2015) usando essa representação para a sumarização multidocumento, e também porque ela apresentou o melhor desempenho nos experimentos realizados e apresentados no Apêndice A. Já a relevância (peso) de um conceito é estimada com base em sua posição, centralidade (em nível de sentenças), e nos pesos de outros conceitos adjacentes. No Apêndice A são apresentados os experimentos realizados comparando diferentes formas de representação e métodos para a ponderação da relevância dos conceitos.

Uma visão geral da abordagem proposta é apresentada na Figura 3, e as sete etapas ilustradas são brevemente descritas a seguir.

1. Pré-processamento: Nesta etapa, o documento de entrada é pré-processado utilizando a ferramenta *Stanford Natural Language Processing Toolkit* (CoreNLP) (MANNING et al., 2014). As tarefas de PLN executadas são: segmentação de sentenças, tokenização, lematização, atribuição das classes gramaticais, análise sintática e resolução de correferência.

2. Extração de Conceitos: Esta etapa é responsável pela extração dos conceitos. Neste trabalho, bigramas são utilizados para representar conceitos. Bigramas formados somente por *stopwords* ou que possuem símbolos de pontuação são removidos (GILLICK et al., 2009; BOUDIN; MOUGARD; FAVRE, 2015).

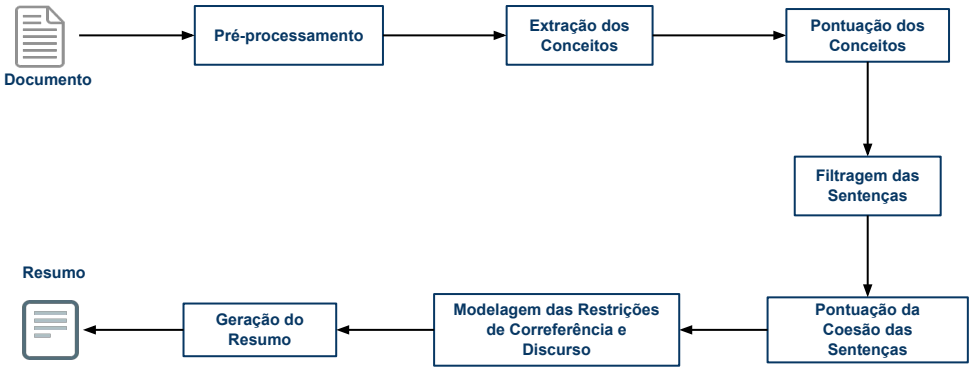


Figura 3 – Visão geral da abordagem proposta.

3. Pontuação dos Conceitos: O método de ponderação proposto é aplicado a cada conceito extraído na etapa anterior para estimar a sua relevância para o documento. Uma das contribuições deste trabalho reside na nova estratégia de atribuição de pesos somente para a primeira ocorrência de um conceito no documento.

4. Pontuação da Coesão Local das Sentenças: Nesta etapa, utilizamos o modelo de Grafo de Entidades (GUINAUDEAU; STRUBE, 2013) para atribuir uma pontuação aproximada relativa à coesão local de cada sentença do documento. Dessa forma, podemos integrá-lo ao modelo de otimização, permitindo a maximização simultânea com a cobertura de conceitos relevantes.

5. Filtragem das Sentenças: Sentenças muito curtas podem não ser representativas o suficiente para contribuir com a informatividade dos resumos, da mesma maneira que frases muito longas podem ser um desperdício de espaço. Desta forma, com base em trabalhos anteriores (BOUDIN; MOUGARD; FAVRE, 2015; FERREIRA et al., 2013), frases com dez ou menos palavras ou com setenta ou mais palavras são removidas. Além disso, sentenças duplicadas também são removidas. Para garantir a integridade do modelo, as remoções feitas nesta etapa são realizadas diretamente no modelo de PLI através da inclusão da restrição $\sum_{s_r \in S_r} s_r \leq 0$, de forma a garantir que todas as sentenças removidas não possam ser selecionadas para o resumo.

6. Modelagem das Restrições de Legibilidade: Nesta etapa, as dependências entre as correferências pronominais e seus antecedentes (ou referentes) e as dependências explícitas de discurso são identificadas e modeladas como restrições no modelo de PLI.

7. Geração do Sumário: Nesta etapa, o processo de geração do resumo é tratado como um problema de otimização, tal como apresentado na Equação 4.1. Para a resolução deste problema, adotamos a ferramenta GNU Linear Programming Kit (GLPK) package[1]. Um valor binário é atribuído para cada sentença do documento,

[1] https://www.gnu.org/software/glpk/

no qual o valor 1 indica que a sentença foi selecionada para o resumo e o valor 0 que ela não foi selecionada. Além disso, uma sentença somente é inserida no resumo caso ela não possua similaridade do cosseno maior que 0,5 com nenhuma outra sentença já incluída no resumo (HONG et al., 2014).

Devido à sua importância para a abordagem proposta, as etapas 3, 4 e 6 são detalhadas nas subseções a seguir.

4.1.1 Ponderação da Relevância dos Conceitos

Diversos métodos para ponderar a importância dos conceitos têm sido propostos e avaliados na literatura. Alguns desses métodos são baseados na frequência de documentos que mencionam o conceito (sumarização multidocumento) (GILLICK et al., 2009; BOUDIN; MOUGARD; FAVRE, 2015) ou aplicando algoritmos de regressão para estimar, a partir de um conjunto de exemplos de treinamento, as pontuações dos conceitos (CAO et al., 2015; LI; LIU; ZHAO, 2015). Contudo, no melhor do conhecimento do autor desta tese, nenhum trabalho encontrado realizou uma avaliação de vários métodos para a ponderação de conceitos na tarefa de sumarização monodocumento. Para suprir essa lacuna, no Apêndice A são apresentados os resultados de diversos experimentos que avaliam diferentes formas de representação para os conceitos, bem como vários métodos para mensurar a sua importância.

Apesar de ser um dos métodos mais simples, a frequência de documentos tem apresentado bons resultados para a sumarização multidocumento (BOUDIN; MOUGARD; FAVRE, 2015; LI; LIU; ZHAO, 2015). Contudo, como este trabalho foca na sumarização monodocumento, o método de frequência dos documentos não é adequado. Dessa forma, a frequência das sentenças que mencionam o conceito foi adotada. O método de posição das sentenças já é tradicionalmente aplicado em trabalhos de sumarização monodocumento, devido aos bons resultados apresentados na literatura (OUYANG et al., 2010; FERREIRA et al., 2013). Com base nos bons resultados obtidos e também pela alta diversidade de resumos gerados por esses dois métodos nos experimentos conduzidos e apresentados no Apêndice A, este trabalho propõe combiná-los visando melhor estimar os pesos dos conceitos extraídos. Na Equação 4.2, é apresentado como é computada a relevância de um conceito usando a combinação proposta.

$$Pontuacao(c_i) = Posicao(c_i) \times FreqSentencas(c_i) \tag{4.2}$$

$$Posicao(c_i) = 1 - \frac{Pos_{s_{c_i}}}{S} \tag{4.3}$$

na qual,

- $Posicao(c_i)$ retorna a posição da sentença na qual c_i está contido. A Equação 4.3 demonstra como essa pontuação é calculada;
- $FreqSentenças(c_i)$ retorna o total de sentenças que mencionam o conceito c_i;
- $Pos_{s_{c_i}}$ retorna o índice da posição da primeira sentença que contém o conceito c_i[2];
- S é o total de sentenças presentes no documento de entrada.

Algoritmos tradicionais baseados em grafos, tais como o TextRank (MIHALCEA; TARAU, 2004) e PageRank (BRIN; PAGE, 1998), adotam a importância dos nós vizinhos para mensurar a relevância de um vértice. Essa ideia é baseada na suposição de que um vértice é importante se ele também está inserido em um contexto relevante. Baseado nessa premissa, o método proposto representa os conceitos extraídos do documento como vértices, e as arestas denotam relações de adjacência entre dois conceitos. Na Equação 4.4 é definido como o peso de um conceito é computado.

$$Peso(c_i) = Pontuacao(c_i) + \frac{\sum_{j \in In(c_i)} Pontuacao(c_j)}{|In(c_i)|} \tag{4.4}$$

na qual,

- $Pontuacao(c_i)$ é a pontuação atribuída a um conceito c_i, conforme calculado na Equação 4.2,
- $In(c_i)$ é o conjunto de vértices incidentes do conceito c_i,
- $|In(c_i)|$ é o total de vértices incidentes do conceito c_i.

Os experimentos relatados no Capítulo 3 demonstram o quão importante o método de posição das sentenças é para a tarefa de pontuação das sentenças, principalmente para a sumarização monodocumento. Por essa razão, este trabalho propõe incorporar esse aspecto diretamente na estratégia de distribuição dos pesos dos conceitos. Para isso, depois de computar os pesos, somente a primeira ocorrência de um conceito no documento recebe a sua pontuação, as demais ocorrências recebem valor zero. A ideia principal dessa estratégia é privilegiar frases no início do documento, se elas possuírem conceitos relevantes, mas também contemplar frases no meio ou no final do documento, caso elas insiram novos conceitos ainda não selecionados para compor o resumo.

4.1.2 Pontuação da Coesão Local das Sentenças

O modelo de Grafo de Entidades (GE) (GUINAUDEAU; STRUBE, 2013) usado neste trabalho é baseado em um outro modelo chamado de Grade de Entidades, proposto por Barzilay e Lapata (2008). O modelo de grade de entidades é uma representação de discurso baseada em entidades inspirada pela Teoria da Centralidade (GROSZ; WEINSTEIN;

[2] O índice das sentenças começa a ser contado pelo número zero.

JOSHI, 1995), usada para mensurar a coesão local de um texto. A coesão local é medida analisando a similaridade entre as sentenças de um texto em relação às frases vizinhas. Dessa forma, supõe-se que sentenças adjacentes muito conectadas apresentam uma boa coesão léxica. A intuição desse modelo é que entidades compartilhadas por sentenças subsequentes contribuem para a coesão local de um texto. A grade de entidades é uma matriz $M_{N\times E}$, na qual as linhas correspondem às S sentenças e as colunas são as E entidades. Para cada entidade e_j e sentença s_i no texto, existe uma entrada na matriz M_{ij} contendo informações indicando a ausência ou presença de e_j em s_i. Se e_j não está presente em s_i, a célula M_{ij} é atualizada com o símbolo "-". Caso contrário, a célula M_{ij} é atualizada com o papel sintático da entidade e_i, que pode ser sujeito ("S"), objeto ("O") ou "X" para qualquer outro papel. Os autores então computam a coesão local de um texto utilizando algoritmos de aprendizagem supervisionados analisando o padrão de transição entre as sentenças adjacentes.

Guinaudeau e Strube (2013) adaptaram o modelo de grade de entidades utilizando um grafo bipartido, chamado de grafo de entidades, e computaram a coesão local de um texto baseados nesse grafo, de forma não supervisionada. O grafo bipartido $G = (V_s, V_e, A, p)$ é composto por dois tipos de vértices independentes que correspondem às sentenças V_s e às entidades V_e, e por um conjunto de arestas A que associam uma sentença a uma entidade com um certo peso p. Os autores adotaram substantivos como entidades. Para calcular a coesão local de um texto, os autores realizam uma projeção de um único modo, no qual os vértices são as sentenças e uma aresta é criada entre uma frase e as orações seguintes se essas compartilham uma mesma entidade. A pontuação da coesão local do texto é dada pela média do número de arestas que saem de cada vértice (sentença) na projeção. Quanto maior a pontuação obtida, mais coeso é o texto. Os autores demonstraram que a abordagem proposta é computacionalmente mais eficiente do que o modelo original de grade de entidades, além de aliviar problemas de dispersão dos dados. Além disso, os autores avaliaram diversos tipos de projeções, sem pesos nas arestas, ou utilizando o papel sintático das entidades para atribuir pesos as arestas, de modo a dar maior pontuação às entidades que desempenham papéis de sujeito ou objeto. Nos experimentos conduzidos em (GUINAUDEAU; STRUBE, 2013), a projeção sem pesos obteve melhor performance do que considerando os pesos das arestas.

Neste trabalho, o modelo de grafo de entidades é utilizado para mensurar a coesão local dos resumos gerados, devido aos bons resultados obtidos por Guinaudeau e Strube (2013) e também pela boa performance obtida pelos recentes trabalhos de Parveen, Ramsl e Strube (2015a), Parveen e Strube (2015b), que utilizaram esse modelo na sumarização monodocumento. Na Figura 4 é apresentado um exemplo de um grafo de entidades criado a partir das sentenças s_1, s_2, s_3 e s_4 a seguir[3]. A projeção de um modo sem pesos desse grafo é ilustrada na Figura 5. Essa projeção mostra as sentenças (vértices), e as arestas

[3] As frases foram extraídas do corpus do DUC 2002.

representam entidades compartilhadas entre uma sentença e suas frases subsequentes. Como os pesos das arestas não foram utilizados neste trabalho, os papéis sintáticos das entidades não são considerados nesta etapa.

> S_1: *Hurricane*$_{[E_1]}$ *Gilbert*$_{[E_2]}$ *slammed into Kingston*$_{[E_3]}$ *on monday*$_{[E_4]}$ *with torrential rains*$_{[E_5]}$.
>
> S_2: *No serious injuries*$_{[E_6]}$ *were immediately reported in Kingston*$_{[E_3]}$.
>
> S_3: *For half an hour*$_{[E_7]}$, *the hurricane*$_{[E_1]}$ *lashed the city*$_{[E_8]}$, *tearing branches*$_{[E_9]}$ *from trees*$_{[E_{10}]}$, *blowing down fences*$_{[E_{11}]}$ *and whipping paper*$_{[E_{12}]}$ *through the air*$_{[E_{13}]}$.
>
> S_4: *The National*$_{[E_{14}]}$ *Weather*$_{[E_{15}]}$ *Service*$_{[E_{16}]}$ *reported heavy damage*$_{[E_{17}]}$ *to Kingston*$_{[E_3]}$*'s airport*$_{[E_{18}]}$ *and aircraft*$_{[E_{19}]}$ *parked on its fields*$_{[E_{20}]}$.

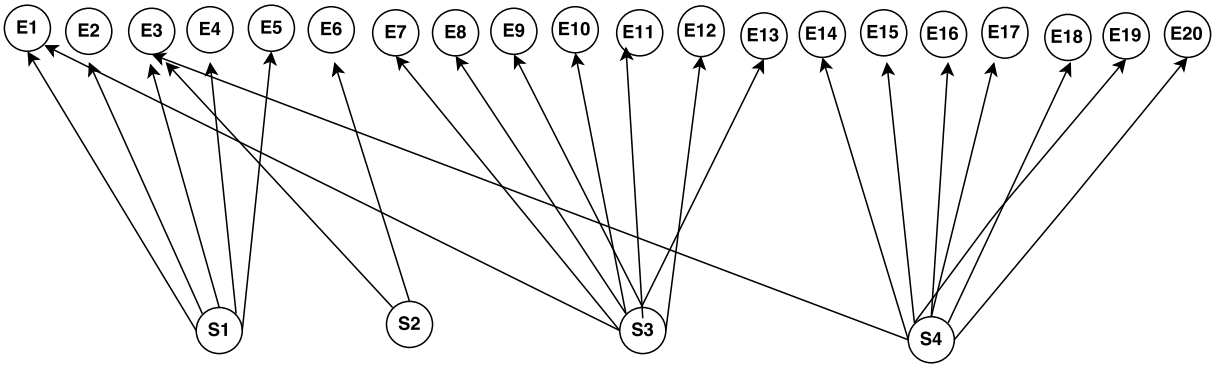


Figura 4 – Exemplo de um grafo de entidades das sentenças S_1 até S_4 listadas anteriormente.

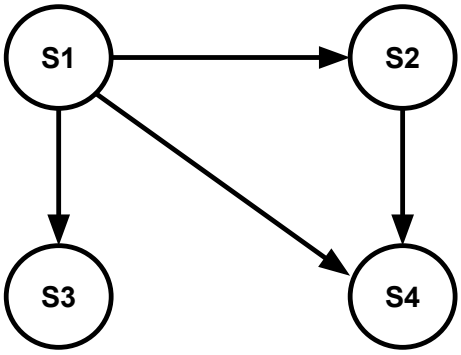


Figura 5 – Projeção de um modo sem pesos do grafo de entidade apresentado na Figura 4.

Para promover a integração da coesão local gerada a partir do grafo de entidades com o modelo de otimização adotado neste trabalho, seguimos a sugestão de Parveen, Ramsl e Strube (2015a), que utilizaram a coesão local de cada sentença presente na projeção criada a partir do grafo de entidades. Dessa forma, é possível incluí-las na função objetivo do modelo de forma a maximizar tanto a cobertura de conceitos relevantes, como também

a seleção de sentenças que maximizem a coesão local do resumo. A pontuação de coesão de uma sentença utilizada neste trabalho é dada pela Equação 4.5.

$$Coesao(s_j) = grau\_saida(s_j, P) \tag{4.5}$$

na qual,

- s_j é uma sentença no documento de entrada,
- P é a projeção de um modo sem pesos do grafo de entidades, e
- $grau\_saida(s_j, P)$ retorna o número de arestas que saem do vértice correspondente a sentença s_j.

4.1.3 Modelagem das Restrições de Correferência e Relações Explícitas de Discurso

Além da coesão local gerada a partir do grafo de entidades criado na etapa anterior, este trabalho explora a inclusão de restrições explícitas para evitar tradicionais problemas de correferências em aberto e quebras no fluxo de discurso entre as sentenças. Para minimizar esses dois problemas, restrições são incluídas no modelo de PLI utilizando **(i)** *Resolução de Correferências* e **(ii)** *Análise Explícita de Discurso*, ambas detalhadas a seguir.

Resolução de Correferência (RC) consiste na identificação de palavras ou expressões que fazem menções a entidades ou conceitos introduzidos anteriormente (LEE et al., 2013). Normalmente, os sistemas de RC fornecem uma cadeia de correferências que contém a entidade referenciada e todas as suas menções ao longo de um documento. RC é uma tarefa complexa e muito desafiadora, de forma que os atuais sistemas do estado da arte ainda são propensos a cometer muitos erros. Por esta razão, este trabalho foca na resolução de correferências pronominais, porque é um tipo de correferência mais simples e menos suscetível a erros. Contudo, as cadeias de correferências do pronome *It* também foram ignoradas, devido ao elevado número de erros relacionados a esse pronome.

Para ilustrar como a abordagem proposta utiliza as restrições de correferências pronominais, considere as frases s_1 e s_2 apresentadas a seguir[4], nas quais o pronome "*She*" presente em s_2 se refere à entidade "Arianna Huffington" introduzida em s_1. Dessa forma, o modelo baseado em PLI proposto inclui uma nova restrição, indicando que s_2 contém uma dependência com s_1, ou seja, para que s_2 seja completamente entendida, é necessário que s_1 também seja inserida no resumo.

> s_1: *As president and editor-in-chief of The Huffington Post, you would expect* ***Arianna Huffington*** *to be living her life at 100 mph.*

[4] As frases foram extraídas do corpus CNN.

> s_2: ***She*** *was listed on Time Magazine's "Time 100" list of the world's 100 most influential people in 2006 and 2011 services.*

Além de modelar as dependências das correferências, restrições de análise de discurso também foram consideradas. Análise de Discurso (PITLER et al., 2008) fornece a base para a representação da coesão do discurso em nível de documento. As relações de discurso podem ser *explícitas*, sendo essas identificadas facilmente pela presença de conectivos de discurso ou marcadores, tais como *but*, *however*; ou *implícitas*, que são mais difíceis de serem reconhecidas, porque são inferidas apenas pelo contexto.

O foco neste trabalho são os marcadores de discurso explícitos, tais como *but*, *however*, *moreover*, *thus*, entre outros, que são usados para indicar dependências de discurso entre duas sentenças adjacentes. Para ilustrar essas dependências, considere as frases s_1 e s_2 apresentadas a seguir[5], que são conectadas pelo marcador de discurso "*but*" em negrito presente na sentença s_2. Para modelar essa dependência, uma nova restrição é adicionada ao modelo de PLI para indicar que s_2 depende de s_1, ou seja, s_2 só é aceita na solução do modelo se s_1 também estiver presente no resumo. Em outras palavras, o marcador "*but*" indica que s_2, para ser plenamente compreendida, precisa de informações que estão presentes na sentença s_1.

> s_1: *Taylor says sellers from 89 countries use fulfillment by Amazon to sell goods to U.S. customers.*
>
> s_2: ***But*** *the advantages Amazon gains by enabling micro-exports extend beyond the fees charged for its services.*

A abordagem proposta estende os trabalhos anteriores na literatura buscando evitar problemas de correferência em aberto e quebra no fluxo de discurso entre sentenças, por meio da inclusão da restrição $Ds_j \leq \sum_d^D s_d$ (Equação 4.1e) adicionada ao modelo de PLI. Essa restrição é usada para representar cada correferência ou relação de discurso como *dependência* entre duas frases. Na verdade, tal restrição garante que se uma sentença s_j tem dependência com outras D sentenças s_d, então s_j só será inserida no resumo se todas as frases de que s_j depende também forem inseridas.

4.1.4 Exemplo de Execução da Abordagem Proposta

Nesta seção é apresentado um exemplo de execução da abordagem proposta para facilitar o entendimento das suas principais etapas. Para isso, utilizaremos como entrada um documento d composto pelas seguintes sentenças[6]:

> S_1: *Hurricane Gilbert swept toward the Dominican Republic Sunday.*

[5] As frases foram extraídas do corpus CNN.

[6] As frases foram extraídas do documento AP880911-0016 do corpus do DUC 2002.

S_2: *"There is no need for alarm" Civil Defense Director Eugenio Cabral said in a television alert.*

S_3: *He said residents of the province of Barahona should closely follow the Hurricane Gilbert movement.*

O Quadro 8 ilustra cada sentença do documento d e os seus respectivos bigramas extraídos na etapa de *Extração de Conceitos*. Os bigramas destacados em itálico são removidos por serem compostos somente por *stopwords*.

Quadro 8 – Sentenças do documento d e seus respectivos conceitos (bigramas) extraídos.

Sentenças	**Conceitos**
S_1	hurricane gilbert - gilbert swept - swept toward - toward the - the dominican - dominican republic - republic sunday
S_2	*there is* - *is no* - no need - need for - for alarm - civil defense - defense director - director eugenio - eugenio cabral - cabral said - said in - a television - television alert
S_3	he said - said residents - residents of - *of the* - the province - province of - of Barahona - barahona should - should closely - closely follow - follow the - the hurricane - hurricane gilbert - gilbert movement

Após a extração de conceitos, cada um deles é analisado para computar sua relevância na etapa de *Pontuação de Conceitos*. A Tabela 11 apresenta cada conceito extraído e seus respectivos pesos (normalizados entre 0 e 1). Vale relembrar que apenas a primeira ocorrência de um conceito recebe seu escore de importância, as demais menções recebem pontuação igual a zero.

Tabela 11 – Conceitos extraídos e seus respectivos pesos.

Conceito	**Peso**
Hurricane Gilbert	1,0
Gilbert swept	0,978
Gilbert movement	0,913
swept toward - toward the - the Dominican - Dominican Republic	0,783
Republic Sunday	0,717
no need - the Hurricane	0,587
need for - for alarm - Civil Defense - Defense Director - Director Eugenio - Eugenio Cabral - Cabral said - said in - a television	0,522
television alert	0,457
He said	0,326
said residents - residents of - the province - province of - of Barahona - Barahona should - should closely - closely follow - follow the	0,261

Para fins de ilustração, a fase de filtragem das sentenças foi desconsiderada. Após a etapa de ponderação dos pesos dos conceitos, inicia-se a computação da coesão local das sentenças usando o modelo de grafo de entidades. Para cada sentença do documento, é gerada uma pontuação que representa o quanto de entidades ela compartilha com sentenças subsequentes. Como o documento d usado neste exemplo só possui três sentenças, e a segunda não possui nenhuma entidade (substantivo) em comum com a terceira frase, s_2 recebe uma pontuação igual a zero. A sentença s_3, por ser a última do documento, também recebe um escore igual a zero. Já a sentença s_1 recebe um escore igual a $0,5$ porque possui entidades compartilhadas com s_3.

Por fim, é realizada a seleção das sentenças para compor o resumo usando o modelo de PLI apresentado na Equação 4.1. Para fins de exemplificação, o limiar L do tamanho máximo do resumo foi definido como $L = 20$, ou seja, o resumo terá no máximo vinte palavras. É possível observar que a sentença s_3 possui uma menção, usando o pronome "He", ao nome "Eugenio Cabral" presente na sentença s_2. Por isso, uma restrição de dependência, indicando que s_3 depende s_2 para ser completamente entendida, é criada conforme apresentada na Equação 4.1e. Com isso, a função objetivo, a restrição do tamanho de resumo, e a restrição de dependência entre s_3 e s_2 presentes no modelo de PLI do exemplo aqui ilustrado são definidos como exemplificado a seguir. Por questões de espaço e para facilitar o entendimento, as restrições (Equações 4.1c e 4.1d) que garantem a consistência do modelo são omitidas.

- $Maximize \quad 1,0 * c_1 + 0,978 * c_2 + 0,913 * c_3 + 0,783 * c_4 + 0,783 * c_5 + 0,783 * c_6 + 0,783 * c_7 + 0,717 * c_8 + 0,587 * c_9 + 0,587 * c_{10} + 0,522 * c_{11} + 0,522 * c_{12} + 0,522 * c_{13} + 0,522 * c_{14} + 0,522 * c_{15} + 0,522 * c_{16} + 0,522 * c_{17} + 0,522 * c_{18} + 0,522 * c_{19} + 0,457 * c_{20} + 0,326 * c_{21} + 0,261 * c_{22} + 0,261 * c_{23} + 0,261 * c_{24} + 0,261 * c_{25} + 0,261 * c_{26} + 0,261 * c_{27} + 0,261 * c_{28} + 0,261 * c_{29} + 0,261 * c_{30} + 0,0 * s_3 + 0,5 * s_1 + 0,0 * s_2$
- $Subject\ To \quad 15 * s_3 + 8 * s_1 + 16 * s_2 <= 20$
- $1 \times s_3 \leq s_2$, ou seja, se s_3 for selecionada para compor o resumo, então s_2 também deve ser selecionada; caso contrário, resultará na restrição inválida $1 \leq 0$.

O resumo gerado no exemplo apresentado contém somente a sentença s_1 "*Hurricane Gilbert swept toward the Dominican Republic Sunday.*", pois ela contempla a maior quantidade de conceitos importantes, maior pontuação de coesão, não possui dependência com nenhuma outra sentença, e satisfaz a restrição do tamanho máximo que o resumo gerado deve ter.

4.2 Experimentos

Nesta seção, são apresentados e discutidos os resultados dos experimentos conduzidos para avaliar diferentes aspectos da abordagem proposta. Três experimentos foram executados abordando as seguintes questões: **(i)** Avaliação dos métodos de ponderação de conceitos isoladamente com e sem a adoção da estratégia de distribuição de pesos proposta (Subseção 4.2.2); **(ii)** Análise do impacto da inclusão das restrições de correferência e análise de discurso, além da pontuação da coesão local das sentenças, em termos das medidas de avaliação do ROUGE (Subseção 4.2.3); e **(iii)** Comparação do desempenho da abordagem proposta com outros sistemas de sumarização monodocumento do estado da arte (Subseção 4.2.4).

4.2.1 Configurações do Experimento

Todos os experimentos realizados utilizaram os corpora do DUC 2001, DUC 2002 e CNN. Na Tabela 12 são apresentadas algumas estatísticas básicas para estes corpora.

Tabela 12 – Estatísticas básicas dos corpora utilizados nos experimentos.

Corpus	#Documentos	#Sentenças	#Palavras
CNN	3.000	115.649	2.628.336
DUC 2001	308	11.026	269.990
DUC 2002	533	14.370	348.012

Para avaliar os resumos gerados, as seguintes medidas foram adotadas:

ROUGE: As medidas de cobertura do ROUGE-1 (R-1) e ROUGE-2 (R-2) foram adotadas em todos os experimentos realizados. A versão 1.5.5 do ROUGE foi empregada com os parâmetros: *-m -f A*. Uma vez que os modelos DUC têm um limiar baseado na contagem de palavras, nesses corpora, o parâmetro *N -l* foi usado para truncar todos os resumos gerados para conter *N* palavras.

Intersecção de Sentenças (IS) (MANI, 2001; FERREIRA et al., 2013): Essa medida mensura a intersecção de sentenças entre o resumo candidato e o resumo de referência disponível. Vale salientar que essa medida só pode ser computada quando resumos de referência extrativos estão disponíveis. Dessa forma, a medida IS é adotada somente nas avaliações conduzidas no corpus CNN.

Os resumos gerados no corpus CNN utilizaram a taxa de compressão de 10% do número de sentenças do documento de entrada. Para os corpora do DUC 2001 e DUC 2002, o limiar dos resumos foi definido para no máximo 105 palavras. Esse limiar foi adotado para gerar resumos com tamanhos comparáveis com os gerados pelos trabalhos relacionados, que possuem aproximadamente 100 palavras.

4.2.2 Avaliação dos Métodos de Ponderação de Conceitos

Este primeiro experimento avalia a performance de cada método de pontuação (posição e frequência das sentenças) adotado para compor o método de ponderação dos conceitos proposto na Equação 4.4. Para isso, a segunda parte da função objetivo apresentada na Equação 4.1, referente à coesão local das sentenças, não foi considerada. Os métodos de pontuação dos conceitos foram avaliados utilizando as estratégias de distribuição de pesos: Primeira Ocorrência (PO) e Todas as Ocorrências (TO). A primeira estratégia é a proposta deste trabalho, que atribui peso somente para a primeira ocorrência de um conceito, enquanto que a segunda estratégia consiste na abordagem tradicional, que atribui peso para todas as ocorrências de um conceito. Na Tabela 13 são apresentados os resultados obtidos neste experimento em termos das medidas de cobertura do ROUGE-1 (R-1) e ROUGE-2 (R-2), e da medida de Intersecção de Sentenças (IS).

A estratégia proposta de atribuir pesos somente à primeira ocorrência de um conceito obteve melhor desempenho do que a distribuição de pesos tradicional em todas as comparações realizadas com base nas três medidas de avaliação adotadas. Em quase todas as comparações, a estratégia PO apresentou melhorias significativas comparando as medidas do R-1 e R-2 ao nível de 95% de confiança aplicando o teste estatístico de *Wilcoxon signed rank* (GIBBONS; CHAKRABORTI, 2003). Em relação aos métodos de ponderação dos conceitos individualmente, o método de posição das sentenças apresentou melhor desempenho do que o método de frequência das sentenças em todos os corpora avaliados. O método combinado proposto obteve resultados melhores do que as respectivas aplicações individuais em todos os cenários.

Na linha (4) da Tabela 13, pode-se notar que, adicionando os pesos dos conceitos adjacentes para a configuração apresentada na linha (3), os resultados apresentaram uma ligeira melhora no desempenho nos corpora do DUC 2002 e CNN, em todas as medidas de avaliação. A única exceção foi uma leve queda no desempenho em termos das medidas do ROUGE no DUC 2001, após a consideração dos pesos dos conceitos vizinhos.

Os resultados obtidos neste experimento demonstram que: **(i)** A estratégia de primeira ocorrência proposta obteve melhorias estatísticas em relação à tradicional estratégia de distribuição dos pesos dos conceitos; **(ii)** A combinação do tradicional método de posição das sentenças com o método de frequência das sentenças obteve um bom desempenho, levando a resultados melhores do que considerá-los isoladamente; e **(iii)** A ideia de considerar os pesos dos conceitos adjacentes não foi tão eficiente quanto esperado, mesmo assim, apresentou resultados melhores em dois dos três corpora avaliados.

Tabela 13 – Resultados (%) e desvio padrão entre parênteses dos métodos de pontuação dos conceitos. O melhor desempenho global de cada corpus é destacado em negrito, e o grupo de configurações estatisticamente semelhantes, se existir, é indicado pelo símbolo †.

Métodos de Pontuação	Estratégia de Distribuição	**CNN**		
		IS	R-1	R-2
1) Frequência das Sentenças	Primeira Ocorrência	26,35	56,22 (20,04)	37,23 (26,21)
	Todas as Ocorrências	21,02	54,10 (19,32)	32,84 (25,25)
2) Posição das Sentenças	Primeira Ocorrência	29,49	56,77 (20,22)	39,70 (25,81)
	Todas as Ocorrências	25,33	56,45 (19,52)	37,21 (25,62)
3) 1 + 2	Primeira Ocorrência	30,94	57,11 (20,30)	40,78† (25,55)
	Todas as Ocorrências	28,53	56,97 (19,90)	39,42 (25,50)
4) 3 + Vizinhos	Primeira Ocorrência	**31,00**	**57,47** (20,07)	**40,98** (25,40)
	Todas as Ocorrências	28,42	56,95 (20,09)	39,40 (25,67)
		DUC 2001		
		IS	R-1	R-2
1) Frequência das Sentenças	Primeira Ocorrência	-	44,59† (9,64)	19,28 (11,33)
	Todas as Ocorrências	-	42,09 (8,68)	15,82 (9,60)
2) Posição das Sentenças	Primeira Ocorrência	-	44,65† (9,54)	19,86† (11,28)
	Todas as Ocorrências	-	43,02 (9,38)	17,50 (10,63)
3) 1 + 2	Primeira Ocorrência	-	**45,12** (9,58)	**20,13** (11,31)
	Todas as Ocorrências	-	44,22 (9,60)	19,33 (11,15)
4) 3 + Vizinhos	Primeira Ocorrência	-	45,00† (9,72)	20,05† (11,42)
	Todas as Ocorrências	-	44,14 (9,68)	19,31 (11,22)
		DUC 2002		
		IS	R-1	R-2
1) Frequência das Sentenças	Primeira Ocorrência	-	48,23 (8,91)	22,52 (10,09)
	Todas as Ocorrências	-	45,69 (8,81)	19,29 (9,58)
2) Posição das Sentenças	Primeira Ocorrência	-	48,29 (9,03)	22,89 (10,15)
	Todas as Ocorrências	-	47,31 (8,86)	21,59 (10,16)
3) 1 + 2	Primeira Ocorrência	-	48,63 (8,91)	23,23† (10,02)
	Todas as Ocorrências	-	47,84 (8,98)	22,32 (9,92)
4) 3 + Vizinhos	Primeira Ocorrência	-	**48,90** (8,50)	**23,42** (9,75)
	Todas as Ocorrências	-	47,83 (8,97)	22,31 (9,91)

4.2.3 Avaliação da Inclusão das Restrições de Dependência e Pontuação da Coesão Local das Sentenças

Este segundo experimento avaliou o impacto da inclusão das restrições de correferência e discurso, e também da pontuação da coesão local das sentenças gerada a partir do grafo de entidades. Quatro configurações foram avaliadas em conjunto com a abordagem proposta: **(i)** Restrições de Correferência; **(ii)** Restrições de Discurso; **(iii)** Inclusão da pontuação da coesão local das sentenças usando o modelo de Grafo de Entidades (GE) na função objetivo do modelo de otimização adotado; e **(iv)** Adotando em conjunto as

restrições de correferência e de discurso, mais a pontuação da coesão local das sentenças. Para validação estatística dos resultados, o teste de *Wilcoxon signed rank* foi aplicado. Na Tabela 14 são apresentados os resultados dos experimentos.

Tabela 14 – Resultados (%) e desvio padrão entre parênteses da abordagem proposta com a inclusão das restrições de coesão com base nas medidas de cobertura do ROUGE-1 (R-1), ROUGE-2 (R-2), e na medida de Intersecção de Sentenças (IS). O melhor desempenho global de cada corpus é destacado em negrito, e o grupo de configurações estatisticamente semelhantes, se existir, é indicado pelo símbolo †.

Configurações	**CNN**		
	IS	R-1	R-2
1. Abordagem Proposta	31,00	57,47† (20,07)	40,98† (25,40)
2. 1 + Relações de Discurso (RD)	30,68	56,98 (20,01)	40,49 (25,19)
3. 1 + Correferência	31,00	57,17 (20,26)	40,83† (25,57)
4. 1 + Grafo de Entidades (GE)	**31,05**	**57,54** (20,09)	**41,08** (25,38)
5. 1 + RD + Correferência + GE	30,83	56,68 (20,21)	40,34 (25,36)
	DUC 2001		
	IS	R-1	R-2
1. Abordagem Proposta	-	45,00† (9,72)	20,05† (11,42)
2. 1 + Relações de Discurso (RD)	-	44,88 (9,71)	19,93 (11,42)
3. 1 + Correferência	-	44,98† (9,64)	20,03† (11,48)
4. 1 + Grafo de Entidades (GE)	-	**45,32** (9,74)	**20,25** (11,52)
5. 1 + RD + Correferência + GE	-	45,07† (9,74)	20,06† (11.58)
	DUC 2002		
	IS	R-1	R-2
1. Abordagem Proposta	-	**48,90** (8,50)	**23,42** (9,75)
2. 1 + Relações de Discurso (RD)	-	48,78† (8,70)	23,36† (9,87)
3. 1 + Correferência	-	48,80† (8,57)	23,34† (9,78)
4. 1 + Grafo de Entidades (GE)	-	48,85† (8,45)	23,30† (9,76)
5. 1 + RD + Correferência + GE	-	48,65 (8,71)	23,17 (9,85)

Analisando os resultados, é possível observar que o desempenho, em termos das medidas do R-1, R-2 e IS, apresentou uma ligeira diminuição quando as restrições de dependências de discurso e correferências foram incluídas isoladamente. Contudo, essa queda no desempenho não representou nenhuma diferença significativa ao nível de 95% de confiança nos corpora do DUC 2001, DUC 2002 e CNN.

Por outro lado, a inclusão das pontuações de coesão local das sentenças geradas usando o modelo de grafo de entidades ocasionou um pequeno aumento na performance de todas as medidas de avaliações usadas nos corpora do DUC 2001 e CNN. Somente no DUC 2002, observou-se uma diminuição nos resultados nas medidas do R-1 e R-2. O único cenário em que houve uma diferença significativa na variação de desempenho foi no corpus do DUC 2001. Nesse corpus, a inclusão da pontuação da coesão local das sentenças melhorou

estatisticamente os resultados da medida do R-1, em relação à configuração original da abordagem proposta.

A inclusão da combinação das restrições de correferência e de discurso com a pontuação da coesão local das sentenças ocasionou uma diminuição na performance da abordagem proposta nos corpora do DUC 2002 e CNN. No corpus CNN, essa deterioração no desempenho, em termos das medidas do ROUGE, foi significativa em relação à versão original da proposta. Já no DUC 2002, a queda de performance não foi considerável. Somente no DUC 2001 ocorreu um ligeiro aumento na performance, contudo essa diferença não foi estatisticamente significante.

Christensen et al. (2013) aplicaram restrições de correferência e quebra de discurso similares às adotadas neste trabalho, só que na tarefa de sumarização multidocumento. Os autores também relataram uma queda no desempenho com base nas medidas do ROUGE após a inclusão das restrições de coesão. Os resultados obtidos neste experimento sugerem que a estratégia de remover frases que causam problemas de coesão é interessante para evitar problemas de coesão nos resumos gerados, mas as frases removidas também podem conter informações importantes que deveriam ter sido incluídas no resumo.

4.2.4 Comparação com outras Abordagens

Este último experimento compara o desempenho da abordagem proposta com os seguintes sistemas: **(i)** Os melhores sistemas participantes das competições do DUC 2001 e 2002 identificados nos experimentos realizados, sendo eles o Sistema T e o Sistema 28, respectivamente; **(ii)** A tradicional *baseline* que consiste na seleção das primeiras n sentenças no corpus CNN (n depende da taxa de compressão adotada), e das primeiras 100 palavras nos corpora do DUC 2001 e 2002, para formar o resumo de saída; e **(iii)** Os sistemas AutoSummarizer (AutoS) (AUTOSUMMARIZER, 2016), Classifier4J (C4J) (LOTHIAN, 2003) e HP-UFPE FS (FERREIRA et al., 2014). Esses três sistemas apresentaram as três melhores performances nos experimentos realizados por Batista et al. (2015). Na Tabela 15 são apresentados os resultados deste experimento em termos das medidas de cobertura do R-1 e R-2, e da medida de IS.

Duas configurações da abordagem proposta foram comparadas, a primeira levando em consideração a pontuação da coesão local das sentenças gerada a partir do grafo de entidade ($Proposta_{GE}$), e a segunda sem levar em consideração essa pontuação. Como o foco desta tese é na informatividade dos resumos gerados, não incluímos os resultados utilizando as restrições de correferências e discurso porque elas deterioraram a performance em termos das medidas de avaliação adotadas.

No corpus do DUC 2001, a abordagem proposta considerando o modelo grafo de entidades ($Proposta_{GE}$) obteve a melhor performance com base na medida do R-1. Em termos da medida do R-2, o melhor desempenho foi apresentado pelo sistema T, mas essa superioridade só foi significativa em relação ao sistema HP-UFPE FS. Nos corpora do

Tabela 15 – Resultados comparativos (%) e desvio padrão entre parênteses em relação a outras abordagens do estado da arte. O melhor desempenho global de cada corpus é destacado em negrito, e o grupo de sistemas estatisticamente semelhantes, se existir, é indicado por †.

Corpus	Sistemas	IS	R-1	R-2
CNN	AutoS	23,16	48,81 (18,70)	32,74 (22,70)
	Baseline	26,74	45,99 (21,77)	33,49 (25,00)
	C4J	23,89	46,63 (20,32)	32,15 (23,13)
	HP-UFPE FS	24,75	50,71 (20,34)	34,58 (24,38)
	Proposta	31,00	57,47† (20,07)	40,98† (25,40)
	$Proposta_{GE}$	**31,05**	**57,54** (20,09)	**41,08** (25,38)
DUC 2001	AutoS	-	41,92 (9,04)	16,63 (9,95)
	Baseline	-	43,75 (10,47)	19,57† (11,64)
	C4J	-	44,44 (9,85)	19,86† (11,34)
	HP-UFPE FS	-	35,91 (11,78)	11,78 (9,78)
	Proposta	-	45,00 (9,72)	20,05† (11,42)
	$Proposta_{GE}$	-	**45,32** (9,74)	20,25† (11,52)
	Sistema T	-	44,53 (9,23)	**20,27** (10,75)
DUC 2002	AutoS	-	43,79 (8,78)	19,17 (9,31)
	Baseline	-	46,94 (9,20)	22,14 (10,01)
	C4J	-	47,09 (8,93)	22,12 (9,87)
	HP-UFPE FS	-	45,70 (9,31)	20,55 (9,88)
	Proposta	-	**48,90** (8,50)	**23,42** (9,75)
	$Proposta_{GE}$	-	48,85† (8,45)	23,30† (9,76)
	Sistema 28	-	48,07 (8,90)	22,88 (9,96)

DUC 2002 e CNN, as duas configurações da abordagem proposta apresentaram resultados estatisticamente superiores a todos os outros sistemas com base na medida do R-1. Em relação à medida IS no corpus CNN, a abordagem proposta ($Proposta_{GE}$) obteve o melhor desempenho.

Nas Tabela 16 e Tabela 17 são apresentados os p-valores obtidos com a aplicação do teste de *Wilcoxon signed rank* comparando as medidas de cobertura do R-1 nos corpora CNN, e DUC 2001-2002, respectivamente. Os p-valores indicando uma diferença estatisticamente significante ao nível de confiança de 95% ou mais são destacados em negrito. Um sinal de maior ($>$) antes do p-valor indica que o sistema na linha obteve maior média na medida de cobertura do R-1 do que o sistema na coluna, enquanto que um sinal negativo ($<$) indica a relação oposta, que o sistema na coluna possui maior R-1 média do que o listado na linha.

Analisando os p-valores, é possível observar que houve uma grande variação no desempenho dos sistemas avaliados em relação à medida de cobertura do R-1, em especial no corpus CNN. Na maioria das comparações realizadas, as duas configurações da abordagem proposta apresentaram melhorias estatisticamente significantes ao nível de 95% ($p - valor < 0.05$) e em alguns casos com superioridade de 99% de confiança

Tabela 16 – P-valores obtidos aplicando o teste de *Wilcoxon signed rank* no corpus CNN.

CNN					
Sistemas	Baseline	C4J	HP-UFPE	Proposta	$Proposta_{GE}$
AutoS	(>) **2,04e-11**	(>) **7,19e-09**	(<) **2,88e-44**	(<) **1,83e-186**	(<) **8,88e-188**
Baseline		(<) **0,2131**	(<) **2,88e-44**	(<) **1,83e-186**	(<) **8,88e-188**
C4J			(<) **1,15e-37**	(<) **6,06e-200**	(<) **2,82e-201**
HP-UFPE				(<) **1,2e-95**	(<) **7,37e-99**
Proposta					(<) 0,064

Tabela 17 – P-valores obtidos aplicando o teste de *Wilcoxon signed rank* nos corpora do DUC 2001-2002.

DUC 2001						
Sistemas	Baseline	C4J	HP-UFPE	Sistema T	Proposta	$Proposta_{GE}$
AutoS	(<) **0,001**	(<) **9,21e-06**	(>) **1,16e-16**	(<) **5,31e-04**	(<) **6,45e-09**	(<) **1,46e-10**
Baseline		(<) **7,57e-04**	(>) **2,12e-24**	(<) 0,618	(<) **5,63e-04**	(<) **9,38e-05**
CJ			(<) **1,22e-28**	(<) 0,500	(<) 0,189	(<) **0,014**
HP-UFPE				(<) **1,43e-19**	(<) **3,25e-31**	(<) **1,73e-32**
Sistema T					(<) 0,156	(<) **0,019**
Proposta						(<) **0,011**
DUC 2002						
Sistemas	Baseline	C4J	HP-UFPE	Sistema 28	Proposta	$Proposta_{GE}$
AutoS	(<) **3,17e-10**	(<) **4,86e-12**	(<) **1,60e-04**	(<) **4,23e-22**	(<) **6,69e-30**	(<) **1,7e-29**
Baseline		(<) 0,469	(>) **2,71e-04**	(<) **7,35e-06**	(<) **3.03e-13**	(<) **5,09e-13**
C4J			(>) **2,37e-05**	**(<) 2,16e-04**	(<) **2,57e-11**	(<) **1,01e-10**
HP-UFPE				(<) **4,516e-12**	(<) **6,396e-20**	(<) **2,22e-19**
Sistema 28					(<) **0,0008**	(<) **0,0011**
Proposta						(>) 0,4583

($p - valor < 0.01$), em relação aos outros sistemas da literatura considerados. Focando nas outras abordagens da literatura avaliadas, os sistemas HP-UFPE, sistema T e sistema 28 nos corpora do CNN, DUC 2001 e DUC 2002, respectivamente, também merecem destaque por obterem um bom desempenho, em muitos casos com superioridade estatística, em relação aos outros sistemas do estado da arte considerados.

Os resultados obtidos demonstram que as duas configurações da abordagem proposta neste capítulo apresentam um desempenho competitivo com os sistemas de sumarização do estado da arte analisados. Além disso, observando o tamanho dos resumos gerados, verificou-se que, em geral, a abordagem proposta utiliza menos palavras para compor os resumos do que os outros sistemas comparados, principalmente nos corpora do DUC. Esse fato sugere que a abordagem proposta consegue maximizar a informatividade dos resumos gerados utilizando uma quantidade menor de palavras do que os demais sistemas avaliados. As estatísticas dos tamanhos em termos de número de sentenças e palavras dos resumos gerados podem ser observadas na Tabela 18.

Na Tabela 19, são apresentados os tempos médios de execução em segundos da abordagem proposta considerando a pontuação de coesão local gerado pelo modelo de grafo de entidades ($Proposta_{GE}$). O tempo de execução foi computado incluindo todas as sete eta-

Tabela 18 – Estatísticas dos tamanhos dos resumos gerados pelos sistemas avaliados com base na quantidade de sentenças e palavras.

Corpus	Sistemas	#Sentenças	#Palavras
CNN	AutoS	10.070	327.327
	Baseline	11.699	296.155
	C4J	11.405	311.473
	HP-UFPE	11.698	394.807
	Proposta	11.693	425.432
	$Proposta_{GE}$	11.693	425.362
DUC 2001	AutoS	1.429	48.597
	Baseline	1.255	35.408
	C4J	1.219	35.854
	HP-UFPE	1.461	34.082
	Proposta	1.287	34.678
	$Proposta_{GE}$	1.300	34.714
	Sistema T	1.230	35.990
DUC 2002	AutoS	1.814	60.841
	Baseline	2.238	60.767
	C4J	2.189	62.158
	HP-UFPE	1.773	64.326
	Proposta	2.244	60.040
	$Proposta_{GE}$	2.263	60.062
	Sistema 28	1.776	61.547

pas do processo de sumarização proposto.[7] A etapa de pré-processamento adotada inclui as tarefas de segmentação de sentenças, tokenização, lematização e atribuição das classes gramaticais. Analisando o tempo médio de execução (em segundos), é possível observar que a abordagem proposta consegue sumarizar um documento em menos de um segundo. Isso demonstra que, além obter resultados competitivos em relação à outras abordagens do estado da arte, a abordagem proposta também requer um baixo custo computacional. Além disso, é possível observar um alto desvio padrão, o que indica que existe uma grande variedade no tamanho dos documentos, o que se reflete diretamente no tempo de geração do resumo.

Tabela 19 – Estatísticas do tempo médio de execução (segundos) e desvio padrão entre parênteses da abordagem proposta usando o modelo de grafo de entidades.

Corpus	Tempo de Execução
CNN	0,068 (0,0795)
DUC 2001	0,094 (0,110)
DUC 2002	0,058 (0,068)

[7] Os tempos de execução foram computados em uma máquina com a seguinte configuração: Intel Core i7-4510U com 2,6 GHz, 16 gigabytes de memória RAM, um terabyte de disco rígido e executando o Windows 8.1 64 bits

4.3 Considerações Finais do Capítulo

Este capítulo apresentou a proposta de uma abordagem baseada em conceitos utilizando programação linear inteira para sumarização monodocumento. A abordagem proposta tem por objetivo maximizar a cobertura de conceitos importantes, bem como evitar redundância e aliviar problemas de coesão dos resumos gerados através de duas estratégias: **(i)** Inclusão de restrições para evitar correferências em aberto e quebra no fluxo de discurso entre sentenças adjacentes, e **(ii)** Integração do modelo de grafo de entidades proposto por Guinaudeau e Strube (2013) para maximizar a coesão local dos resumos. Além disso, uma nova estratégia foi proposta para a distribuição dos pesos dos conceitos que prioriza sentenças no início dos documentos caso elas possuam conceitos importantes, além de sentenças no meio ou no fim dos documentos caso elas introduzam novos conceitos ao resumo gerado.

Os resultados experimentais, em termos das medidas do ROUGE e da medida de intersecção de sentenças, demonstram que a abordagem proposta é viável e apresentou resultados competitivos com diversos sistemas do estado da arte. Além disso, a estratégia de distribuição de pesos proposta apresentou resultados superiores em todos os cenários considerados em relação à tradicional estratégia de pontuar todas as ocorrências dos conceitos. A inclusão de restrições para evitar problemas de correferência em aberto e quebras nas relações de discurso ocasionou uma ligeira queda na performance em termos das medidas de avaliação consideradas, mas essa diminuição não foi significativa na maioria dos cenários. Já a integração com modelo de grafo de entidades (GUINAUDEAU; STRUBE, 2013) melhorou os resultados obtidos pela abordagem proposta nos corpora do DUC 2001 e CNN. Apenas no corpus do DUC 2002 que ocorreu uma queda nos resultados com base nas medidas de avaliações adotadas, mas essa diminuição não foi significativa.

Analisando os resultados obtidos neste capítulo, é possível observar o mesmo padrão visto no Capítulo 3, ou seja, um alto desvio padrão nos resultados com base nas medidas do ROUGE-1 e ROUGE-2, principalmente no corpus CNN. Esse comportamento também foi observado nos experimentos descritos no Apêndice A. Isso sugere que a abordagem proposta nem sempre consegue manter um alto desempenho para todos os documentos de entrada.

5 UMA ABORDAGEM BASEADA EM CONCEITOS UTILIZANDO PLI PARA A SUMARIZAÇÃO MULTIDOCUMENTO

A *Web* proporciona acesso a um volume sem precedentes de informações nos mais diversos formatos, sobre uma grande variedade de tópicos, com vários níveis de confiabilidade e com uma significativa redundância de informações entre diversas fontes. Neste contexto, identificar e filtrar informações de interessante em muitos cenários é uma tarefa complexa. A SAT multidocumento visa gerar um resumo contendo as informações mais relevante a partir de uma coleção de documentos relacionados sobre um mesmo assunto ou evento. Sistemas de sumarização multidocumento podem auxiliar as pessoas, reduzindo o tempo gasto para identificar informações importantes a partir de uma coleção de documentos textuais. Além disso, pelo fato de comparar diferentes documentos, pode-se aumentar a confiabilidade das informações fornecidas no resumo.

Devido a esses aspectos, a SAT multidocumento vem ganhando destaque nos últimos anos, e várias abordagens têm sido propostas e avaliadas (NENKOVA; MCKEOWN, 2012; SAGGION; POIBEAU, 2013; TORRES-MORENO, 2014; GAMBHIR; GUPTA, 2016). Geralmente, os métodos de sumarização multidocumento extrativos envolvem tarefas como o pré-processamento do conjunto de documentos de entrada, a extração e a ponderação da relevância de fragmentos textuais (sentenças ou n-gramas), a eliminação de redundância e algoritmos de ordenação de sentenças (NENKOVA; MCKEOWN, 2012), entre outras.

Na sumarização multidocumento, o grau de redundância das informações entre os documentos é muito alto. Evitar a inclusão de uma ou mais sentenças com uma alta taxa de sobreposição de informações no resumo gerado é essencial para prevenir o desperdício de espaço. Além disso, como as sentenças selecionadas para compor o resumo gerado podem vir de diferentes documentos, a tarefa de ordená-las de forma lógica e coerente é um outro desafio que precisa ser resolvido (BOLLEGALA; OKAZAKI; ISHIZUKA, 2012).

Os resultados experimentais apresentados e discutidos no Capítulo 3 demonstraram que selecionar sentenças com maior pontuação nos seguintes métodos de ponderação resultou, em geral, na geração de resumos informativos com base nas medidas de cobertura do ROUGE-1 e ROUGE-2: **(i) Posição** - sentenças no início dos documentos; e **(ii) Centralidade** - sentenças com mais informações compartilhadas em vários documentos da coleção. Métodos relacionados a posição e centralidade das informações têm sido amplamente investigados para SAT (OUYANG et al., 2010; FERREIRA et al., 2013), sendo explorados de forma individual ou adotando estratégias de combinação.

Ainda no Capítulo 3, observou-se um bom desempenho obtido pelos sistemas basea-

dos em conceitos ICSISumm (GILLICK et al., 2009) e Sume (BOUDIN; MOUGARD; FAVRE, 2015). Abordagens baseadas em conceitos têm sido muito exploradas para a tarefa de sumarização multidocumento nos últimos anos (GILLICK et al., 2009; LI; QIAN; LIU, 2013; BOUDIN; MOUGARD; FAVRE, 2015; LI; LIU; ZHAO, 2015). Duas questões fundamentais que impactam diretamente no desempenho desse tipo de abordagem são a definição de como representar a noção de conceito e, posteriormente, como mensurar a sua relevância. Para representar os conceitos, a maioria dos trabalhos na literatura adotam unidades textuais, como unigramas (CAO et al., 2015; WAN et al., 2015) ou bigramas (GILLICK et al., 2009; LI; QIAN; LIU, 2013; BOUDIN; MOUGARD; FAVRE, 2015; LI; LIU; ZHAO, 2015). A ponderação da importância dos conceitos extraídos é, em geral, mensurada usando métodos individuais, como a quantidade de documentos em que o conceito é mencionado (frequência dos documentos) (GILLICK et al., 2009; BOUDIN; MOUGARD; FAVRE, 2015), ou explorando algoritmos de regressão para estimar a relevância dos conceitos combinando diferentes métodos, como posição e frequência (LI; QIAN; LIU, 2013; LI; LIU; ZHAO, 2015).[1]

Os sistemas que adotam somente o método de frequência dos documentos, negligenciam a importância que outros aspectos, como a posição das sentenças, têm para identificar informações relevantes. Por outro lado, as abordagens supervisionados requerem a disponibilidade de corpora de treinamento, e demandam um custo adicional de anotação dos escores da pontuação da relevância dos conceitos. Neste contexto, ainda é pouco explorada a combinação de diferentes métodos de ponderação de conceitos de forma não supervisionada.

Neste capítulo, apresenta-se uma abordagem baseada em conceitos utilizando PLI para a sumarização multidocumento de artigos de notícias que explora a combinação dos métodos de centralidade e posição. Tal combinação é adotada para filtrar sentenças com um baixo grau de centralidade, e identificar os conceitos mais relevantes para compor o resumo a ser gerado. A solução proposta parte da premissa de que os conceitos mais importantes de um grupo de documentos de notícias ocorrem nas primeiras frases de vários desses documentos, e também são citados em seus títulos. O método proposto baseia-se na estratégia introduzida por Banerjee, Mitra e Sugiyama (2015b) de selecionar como central o documento que compartilha mais informações com outros documentos da coleção, usando-o para orientar o processo de agrupamento das sentenças dos documentos de entrada. Essa estratégia é empregada para auxiliar o processo filtragem das sentenças com um baixo grau de centralidade, e também auxiliam na ordenação das frases no resumo gerado.

As principais contribuições deste capítulo são:

(i) Uma abordagem baseada em conceitos usando PLI que explora os aspectos de centralidade e posição de forma não supervisionada para mensurar a importância dos

[1] É preciso ressaltar que os trabalhos (LI; QIAN; LIU, 2013; LI; LIU; ZHAO, 2015) focam exclusivamente na tarefa de sumarização baseada em tópicos.

conceitos e filtrar sentenças com base na centralidade de seu conteúdo;

(ii) Um método de ponderação de conceitos que combina a centralidade, a posição e também explora os títulos dos documentos para melhor estimar a relevância dos conceitos extraídos;

(iii) Uma estratégia de filtragem baseada no percentual de documentos de entrada que remove grupos de sentenças com poucos membros. Tal estratégia demonstrou ser eficaz para aumentar a informatividade dos resumos gerados e também para reduzir o tempo de execução da abordagem proposta;

(iv) Uma estratégia para a apresentação das sentenças no resumo gerado que utiliza a ordem de aparição das frases no seu documento de origem e também explora informações do processo de agrupamento das sentenças para resolver conflitos que podem ocorrer durante a etapa de ordenação.

O restante deste capítulo é organizado da seguinte forma. Na Seção 5.1 é apresenta a abordagem proposta. Os resultados dos experimentos realizados são discutidos na Seção 5.2. Finalmente, na Seção 5.3 são delineadas as conclusões deste capítulo.

5.1 Abordagem Proposta

A abordagem proposta é baseada na hipótese de que conceitos importantes em artigos de notícias aparecem nas primeiras sentenças da maioria dos documentos de entrada e também em seus títulos. Além disso, a centralidade das informações das sentenças desempenha um papel crucial para a remoção das frases menos relevantes. Assim, dado um grupo de documentos textuais $D = \{d_1, d_2, d_3, ..., d_N\}$, em que cada documento $d_i \in D$ é pré-processado e dividido em um conjunto de sentenças $S_{d_i} = \{s_1^{d_i}, s_2^{d_i}, s_3^{d_i}, ..., s_M^{d_i}\}$, com cada frase contendo uma coleção de conceitos $C_{s_i} = \{c_1, c_2, c_3, ..., c_k\}$, o objetivo da abordagem proposta é gerar um resumo contendo o subconjunto de sentenças $\{s_1^{d_i}, s_1^{d_j}, ..., s_k^{d_N}\}$ que maximiza a cobertura de conceitos relevantes, minimizando a redundância entre as frases selecionadas e respeitando o tamanho máximo do resumo desejado.

A Figura 6 apresenta uma visão geral das principais etapas da abordagem proposta, que são descritas brevemente a seguir.

1. **Pré-processamento:** Cada documento textual $d_i \in D$ é pré-processado adotando a ferramenta *Stanford Natural Language Processing Toolkit* (CoreNLP) (MANNING et al., 2014), executando as seguintes etapas de PLN: segmentação das sentenças, tokenização, lematização e atribuição das classes gramaticais.

2. **Extração e Ponderação dos Conceitos:** Nesta etapa, bigramas são extraídos como uma representação aproximada de conceitos. Posteriormente, o método de pondera-

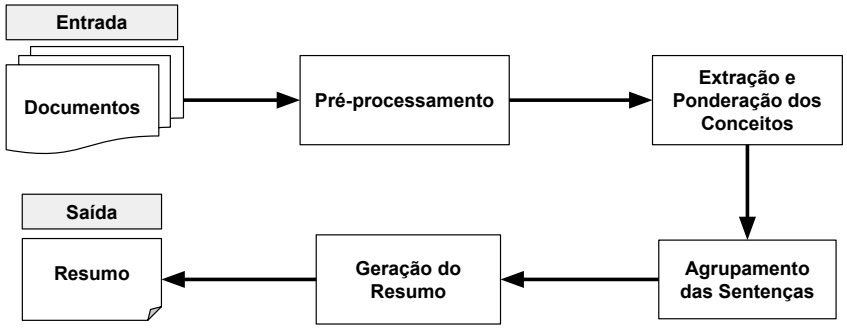


Figura 6 – Visão geral da abordagem proposta.

ção de conceitos proposto é aplicado para mensurar a importância de cada bigrama extraído.

3. **Agrupamento das Sentenças:** Esta etapa identifica o documento com maior similaridade do cosseno com os outros documentos da coleção. Em seguida, esse documento é usado para guiar o processo de agrupamento das sentenças.

4. **Geração de Resumo:** Nesta etapa, a tarefa de seleção das sentenças para compor o resumo gerado é tratada como um problema de otimização assim como proposto por Gillick et al. (2009). PLI é usada para resolver o problema de otimização selecionando as sentenças que maximizam a cobertura dos conceitos relevantes, respeitando o tamanho máximo do resumo a ser gerado. As informações dos grupos de sentenças gerados na etapa anterior são adotadas aqui para **(i)** remover grupos com menos sentenças do que um dado limiar de entrada; e **(ii)** auxiliar o processo de ordenação das sentenças.

Mais detalhes sobre as etapas de *Extração e Ponderação dos Conceitos*, *Agrupamento das Sentenças* e *Geração do Resumo* são apresentadas nas próximas subseções.

5.1.1 Extração e Ponderação dos Conceitos

Escolher uma forma de representação adequada para a noção de conceitos e um método para estimar a relevância desses conceitos são duas questões fundamentais a serem definidas. A maioria das trabalhos na literatura adotam unigramas (CAO et al., 2015) ou bigramas (GILLICK et al., 2009; BOUDIN; MOUGARD; FAVRE, 2015) para representar a noção de conceitos presentes nos documentos de entrada. Este trabalho adota bigramas como conceitos devido ao bom desempenho obtido utilizando essa representação nas medidas de cobertura do ROUGE na literatura e também nos experimentos realizados e descritos na Seção 5.2.2. Como sugerido por Gillick et al. (2009), bigramas formados apenas por *stopwords* são removidos para reduzir a dimensionalidade do conjunto de conceitos extraído.

Em geral, a importância dos conceitos é mensurada na literatura adotando o método que computa a frequência dos documentos que mencionam o conceito (GILLICK et al., 2009; BOUDIN; MOUGARD; FAVRE, 2015), ou combinando vários métodos aplicando algoritmos de aprendizagem supervisionados para combinação de vários métodos (LI; QIAN; LIU, 2013; CAO et al., 2015). Este trabalho propõe um novo método não supervisionado para a ponderação dos conceitos que combina os métodos de frequência dos documentos, posição das sentenças e frequência nos títulos dos documentos. A hipótese é de que um conceito no cenário multidocumento, para ser considerado relevante, deve ser mencionado na maioria dos documentos de entrada, deve ocorrer pela primeira vez no início desses documentos e também deve ser citado em seus títulos. Se um documento não possuir um título, a primeira frase do documento é usada como tal. A Equação 5.1 apresenta como o peso de um conceito é calculado.

$$Peso(c_i) = Pontuacao(c_i) + FreqTitulo(c_i) \tag{5.1}$$

$$Pontuacao(c_i) = FreqDoc(c_i) \times \sum_{d_i \in D} 1 - \frac{Indice_{s_{c_i}}}{S_{d_i}} \tag{5.2}$$

na qual,

- $Pontuacao(c_i)$ é a pontuação de um conceito c_i, calculada conforme apresentado na Equação 5.2;
- $Indice_{s_{c_i}}$ retorna o índice da primeira frase que contém o conceito c_i em cada documento $d_i \in D$;
- S_{d_i} é o número total de frases no documento d_i;
- $FreqDoc(c_i)$ é a quantidade de documentos que mencionam o conceito c_i;
- $FreqTitulo$ é o número total de títulos que contém o conceito c_i.

No final desta etapa, cada frase $s_i \in d_i$ é representada por uma lista de conceitos e seus respectivos pesos. Sentenças contendo apenas conceitos formados por *stopwords* são removidas.

5.1.2 Agrupamento das Sentenças

Dado um conjunto de documentos D, esta etapa executa o processo de agrupamento de todas as frases presentes em D em k grupos, com base nas suas similaridades. Encontrar um valor adequado para k na tarefa de agrupamento é sempre um problema a ser resolvido. Para solucionar esta questão, a estratégia proposta por Banerjee, Mitra e Sugiyama (2015b) de selecionar o documento com o maior índice de centralidade para orientar o processo de agrupamento das frases é adotada aqui. A seleção do documento central d_c é

realizada calculando a similaridade do cosseno de cada documento $d_i \in D$ com todos os documentos de entrada. O documento com maior pontuação média de semelhança é escolhido como central d_c. Tal estratégia de agrupamento apresentou o melhor desempenho nos experimentos relatados por Banerjee, Mitra e Sugiyama (2015b). O conteúdo de todos os documentos é pré-processado para remoção de *stopword* e lemmatização de palavras. O Algoritmo 1 resume a estratégia de identificação do documento central.

Algoritmo 1: Identificação do documento central.

Entrada: O conjunto de documentos D.
Saída: O documento central d_c.

```
1  Início
2      maiorSim = 0
3      para todo d_i ∈ D faça
4          outrosDoc = {}
5          para todo d_j ∈ D faça
6              se i ≠ j então
7                  contOutrosDoc = contOutrosDoc ∪ d_j
8          sim = computarSimCosseno(d_i, outrosDoc)
9          se sim > maiorSim então
10             maiorSim = sim
11             d_c = d_i
```

O Algoritmo 2 apresenta a estratégia adotada para agrupar as sentenças usando o documento central d_c como base para guiar a tarefa. O processo de agrupamento começa utilizando cada frase do documento central d_c para inicializar um novo grupo $g_i \in G$, por exemplo, se d_c possui M sentenças, portanto, M grupos são criados. Cada frase dos outros documentos $d_i \neq d_c$ do conjunto D será atribuída a um grupo com base na sua média de semelhança com os membros do grupo. Esta estratégia calcula a similaridade do cosseno de uma frase $s_j \in d_i$ com cada membro dos grupos $g_i \in G$. Uma sentença s_j é inserida no grupo cujos membros ela possui maior similaridade média, e essa pontuação de semelhança é maior do que um determinado limiar λ, caso contrário s_j é descartada. Esta estratégia pressupõe que frases cujas informações não são mencionadas em vários documentos do grupo não são relevantes o suficiente para serem incluídas no resumo gerado.

Ao final desta etapa, todas as frases da coleção de documentos D estarão agrupadas em M grupos de sentenças. Esses grupos são usados na próxima etapa para filtrar grupos com poucos membros (sentenças) e também para auxiliar o processo de ordenação das sentenças no resumo gerado.

5.1.3 Geração do Resumo

Esta última etapa é responsável pelo processo de seleção das sentenças para compor o resumo a ser gerado. Para isso, esse processo de seleção é tratado como um problema

Algoritmo 2: Agrupamento das sentenças a partir do documento central d_c.

Entrada: O conjunto de documentos D.
Entrada: O documento central d_c.
Saída: Os grupos de sentenças G.

Início
  $G = \{\}$
  $S_c = d_c.sentencas$
  para todo $s_i \in S_c$ **faça**
    $g_i = \{\}$
    $add(s_i, g_i)$
    $G = G \cup g_i$
  $S_{od} = \{\}$
  para todo $d_i \in D$ **faça**
    se $d_i \neq d_c$ **então**
      $S_{od} = S_{od} \cup d_i.sentencas$
  para todo $s_j \in S_{od}$ **faça**
    $maiorMediaSim = 0$
    $g_{similar} = null$
    para todo $g_i \in G$ **faça**
      $S_g = g_i.sentencas$
      $mediaSim = 0$
      para todo $s_g \in S_g$ **faça**
        $mediaSim = mediaSim + computarSimCosseno(s_j, s_g)$
      $mediaSim = \frac{mediaSim}{S_g.tamanho}$
      se $mediaSim > maiorMediaSim$ **então**
        $maiorMediaSim = mediaSim$
        $g_{similar} = g_i$
    se $maiorMediaSim > \lambda$ **então**
      $add(s_j, g_{similar})$

de otimização, adotando o modelo de PLI baseado em conceitos proposto por Gillick et al. (2009).

Primeiro, vislumbrando aumentar a informatividade do resumo gerado e também reduzir o tempo de execução do modelo de PLI, algumas frases isoladas e grupos são removidos com base em seu tamanho. Sentenças com menos de dez ou mais de setenta palavras são removidas com base no pressuposto de que frases muito curtas podem não ser representativas o suficiente para serem incluídas em um resumo, e sentenças muito longas serem um desperdício de espaço (BOUDIN; MOUGARD; FAVRE, 2015). Com base nessa ideia, grupos contendo menos frases do que um determinado limiar de entrada γ também são removidos. A suposição é que se um grupo contém poucas sentenças, isso indica que as informações apresentadas nesse grupo possuem um baixo grau de centralidade. Portanto, elas não são relevantes o suficiente para serem incluídas no resumo gerado.

Após as filtragens, as sentenças não removidas são usadas para construir o modelo de PLI baseado em conceitos conforme apresentado na Equação 5.3. Nessa equação, w_i representa o peso do conceito c_i calculado conforme apresentado na Equação 5.1. A variável binária Occ_{ij} indica a ocorrência de um conceito c_i na frase s_j. A variável l_j representa o

comprimento da sentença s_j e L é o tamanho máximo do resumo a ser gerado. As Equações 5.3c e 5.3d são restrições que garantem a consistência do modelo, assegurando que, se uma sentença for escolhida, todos os seus conceitos também devem ser selecionados, e um conceito só é escolhido se estiver presente em pelo menos uma frase selecionada.

$$max \quad \sum_i w_i c_i \tag{5.3a}$$

$$s.t. \quad \sum_j l_j s_j \leq L \tag{5.3b}$$

$$s_j Occ_{ij} \leq c_i \quad \forall i, j \tag{5.3c}$$

$$\sum_j s_j Occ_{ij} \geq c_i \quad \forall i, j \tag{5.3d}$$

$$c_i, s_j, Occ_{ij} \in \{0, 1\} \quad \forall i, j \tag{5.3e}$$

Este trabalho adota a implementação fornecida pela ferramenta GNU Linear Programming Kit (GLPK)[2] para resolver o problema de otimização. Um valor binário é atribuído a cada sentença, no qual o valor 1 indica que a frase é selecionada para o resumo e 0, o contrário.

O modelo baseado em conceitos usando PLI apresentado na Equação 5.3 consegue simultaneamente: **(i)** maximizar a informatividade do resumo, selecionando conceitos com maior pontuação de relevância; e **(ii)** minimizar redundância, já que uma nova sentença só é selecionada caso acrescente novos conceitos ao resumo gerado. Além disso, uma nova sentença só é inserida no resumo caso ela não possua similaridade do cosseno maior que 0,5 com nenhuma outra sentença já presente no resumo (HONG et al., 2014).

Uma questão final a ser resolvida é como ordenar as frases no resumo gerado. Este trabalho propõe adotar a ordem em que as frases selecionadas aparecem nos seus respectivos documentos de origem para ordená-las no resumo da saída. No entanto, duas frases $s_1^{d_i}$ e $s_1^{d_j}$ com o mesmo índice, mas de dois documentos distintos $d_i \neq d_j$ podem ser selecionadas para compor o resumo. Este problema é resolvido aqui usando os índices dos grupos de sentenças aos quais $s_1^{d_i}$ e $s_1^{d_j}$ pertencem para definir a ordenação final. Como os grupos de sentenças são criados seguindo a ordem das sentenças do documento central que os inicializaram, a ideia é que manter essa ordem é uma estratégia adequada para o resumo gerado. Se ainda assim persistir um impasse entre duas ou mais sentenças, elas são ordenadas com base no índice dos seus documentos. Por exemplo, supondo que as seguintes sentenças foram selecionadas para compor o resumo: $s_1^{d_1}, s_1^{d_2} \in g_1$, $s_3^{d_1} \in g_2$ e $s_2^{d_4} \in g_4$, após o processo de ordenação, a ordem das sentenças no resumo gerado é $s_1^{d_1}$, $s_1^{d_2}$, $s_2^{d_4}$ e $s_3^{d_1}$.

[2] https://www.gnu.org/software/glpk/

5.1.4 Exemplo de Execução da Abordagem Proposta

Nesta seção, apresentamos um exemplo de execução da abordagem proposta, adotando uma configuração padrão definindo os limiares de entrada da similaridade mínima entre as sentenças e do tamanho mínimo do grupo de sentenças como $\lambda = 0$ e $\gamma = 0$, ou seja, nenhuma sentença ou grupo será removida durante as etapas de agrupamento e geração do resumo. O exemplo apresentado a seguir foi executado recebendo como entrada uma coleção de documentos D composta por três documentos d_1, d_2 e d_3, conforme apresentado no Quadro 9.

Quadro 9 – Coleção de documentos contendo três documentos do grupo $d061$ do corpus do DUC 2002.

Documentos	**Sentenças**
d_1	**s1:** *Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast.* **s2:** *The storm was approaching from the southeast with sustained winds of 75 mph.*
d_2	**s1:** *Hurricane Gilbert swept toward Jamaica yesterday with 100-mile-an-hour winds and officials issued warnings to residents on the southern coasts of the Dominican Republic, Haiti and Cuba.* **s2:** *The storm ripped the roofs off houses and caused coastal flooding in Puerto Rico.*
d_3	**s1:** *Hurricane Gilbert slammed into Kingston on Monday with torrential rains and 115 mph winds that ripped roofs off homes and buildings, uprooted trees and downed power lines.* **s2:** *No serious injuries were immediately reported in the city of 750,000 people, which was hit by the full force of the hurricane around noon.*

Após o pré-processamento dos documentos, realiza-se a etapa de extração e ponderação dos conceitos. Para facilitar a apresentação, no Quadro 10 são apresentados os documentos e seus respectivos conceitos (bigramas) extraídos.

Quadro 10 – Documentos $d_i \in D$ e seus respectivos conceitos (bigramas) extraídos.

Documentos	**Conceitos**
d_1	Hurricane Gilbert - Gilbert swept - swept toward - toward the - the Dominican - Dominican Republic - Republic Sunday - the Civil - Civil Defense - Defense alerted - alerted its - its heavily - heavily populated - populated south - south coast - The storm - storm was - was approaching - approaching from - the southeast - southeast with - with sustained - sustained winds - winds of - of 75 - 75 mph
d_2	Hurricane Gilbert - Gilbert swept - swept toward - toward Jamaica - Jamaica yesterday - yesterday with - with 100-mile-an-hour - 100-mile-an-hour winds - winds and - and officials - officials issued - issued warnings - warnings to - to residents - residents on - the southern - southern coasts - coasts of - the Dominican - Dominican Republic - Haiti and - and Cuba - The storm - storm ripped - ripped the - the roofs - roofs off - off houses - houses and - and caused - caused coastal - coastal flooding - flooding in - in Puerto - Puerto Rico
d_3	Hurricane Gilbert - Gilbert slammed - slammed into - into Kingston - Kingston on - on Monday - Monday with - with torrential - torrential rains - rains and - and 115 - 115 mph - mph winds - winds that - that ripped - ripped roofs - roofs off - off homes - homes and - and buildings - uprooted trees - trees and - downed power - power lines - No serious - serious injuries - injuries were - were immediately - immediately reported - reported in - the city - city of - of 750,000 - 750,000 people - was hit - hit by - the full - full force - force of - the hurricane - hurricane around - around noon

Após a extração dos conceitos, cada um deles é avaliado para mensurar a sua importância utilizando o método de ponderação proposto (Equação 5.1). Na Tabela 20 são

apresentados em ordem decrescente os conceitos e seus respectivos pesos. Para facilitar o entendimento, conceitos com o mesmo peso são agrupado em uma mesma linha da Tabela. Não houve uma grande diversidade nos valores dos pesos porque o grupo de documentos D usado neste exemplo possui apenas três documentos, e cada um deles contém apenas duas sentenças. A primeira sentença de cada documento foi usada como título neste exemplo.

Tabela 20 – Conceitos extraídos do grupo de documentos D e seus respectivos pesos.

Conceito	**Peso**
Hurricane Gilbert	1,0
Gilbert swept - swept toward - the Dominican - Dominican Republic	0,47
roofs off	0,33
The storm	0,2
Gilbert slammed - slammed into - into Kingston - toward the - toward Jamaica - Kingston on - Jamaica yesterday - on Monday - yesterday with - with 100-mile-an-hour - Republic Sunday - Monday with - the Civil - 100-mile-an-hour winds - with torrential - Civil Defense - torrential rains - winds and - and officials - Defense alerted - rains and - alerted its - and 115 - officials issued - 115 mph - its heavily - issued warnings - warnings to - mph winds - heavily populated - to residents - winds that - populated south - south coast - that ripped - residents on - the southern - ripped roofs - southern coasts - coasts of - off homes - homes and - and buildings - uprooted trees - Haiti and - and Cuba - trees and - downed power - power lines	0,13
No serious - serious injuries - storm was - storm ripped - was approaching - ripped the - injuries were - were immediately - the roofs - approaching from - the southeast - immediately reported - southeast with - reported in - off houses - the city - houses and - with sustained - city of - and caused - sustained winds - caused coastal - winds of - of 750,000 - 750,000 people - of 75 - coastal flooding - flooding in - was hit - 75 mph - hit by - in Puerto - the full - Puerto Rico - full force - force of - the hurricane - hurricane around - around noon	0,05

Após a etapa de extração e ponderação dos conceitos, as sentenças dos documentos $d_i \in D$ são agrupadas em diferentes grupos. Primeiramente, o documento d_2 foi selecionado como central pelo algoritmo de agrupamento. Dessa forma, as sentenças são agrupadas em dois grupos, conforme apresentado no Quadro 11.

Quadro 11 – Grupos de sentenças e seus respectivos membros.

Grupo	**Sentenças**
g_1	$s_1^{d_1}$, $s_1^{d_2}$, $s_2^{d_3}$
g_2	$s_2^{d_1}$, $s_2^{d_2}$, $s_1^{d_3}$

Por fim, realiza-se a seleção das sentenças para compor o resumo usando o modelo de PLI apresentado na Equação 5.3. Para fins de exemplificação, o limiar L do tamanho máximo do resumo foi definido como $L = 45$, ou seja, o resumo terá no máximo quarenta e cinco palavras. A seguir são apresentadas a função objetivo do modelo de PLI e a restrição do tamanho máximo do resumo a ser gerado.

- $Maximize\ 1,0{*}c_1{+}0,47{*}c_2{+}0,47{*}c_3{+}0,47{*}c_4{+}0,47{*}c_5{+}0,33{*}c_6{+}0,2{*}c_7{+}0,13{*}c_8{+}0,13{*}c_9{+}0,13{*}c_{10}{+}0,13{*}c_{11}{+}0,13{*}c_{12}{+}0,13{*}c_{13}{+}0,13{*}c_{14}{+}0,13{*}c_{15}{+}0,13{*}c_{16}{+}0,13{*}c_{17}{+}0,13{*}c_{18}{+}0,13{*}c_{19}{+}0,13{*}c_{20}{+}0,13{*}c_{21}{+}0,13{*}c_{22}{+}0,13{*}c_{23}{+}0,13{*}c_{24}{+}0,13{*}c_{25}{+}0,13{*}c_{26}{+}0,13{*}c_{27}{+}0,13{*}c_{28}{+}0,13{*}c_{29}{+}0,13{*}c_{30}{+}0,13{*}c_{31}{+}0,13{*}$

$c_{32}+0,13*c_{33}+0,13*c_{34}+0,13*c_{35}+0,13*c_{36}+0,13*c_{37}+0,13*c_{38}+0,13*c_{39}+0,13*c_{40}+0,13*c_{41}+0,13*c_{42}+0,13*c_{43}+0,13*c_{44}+0,13*c_{45}+0,13*c_{46}+0,13*c_{47}+0,13*c_{48}+0,13*c_{49}+0,13*c_{50}+0,13*c_{51}+0,13*c_{52}+0,13*c_{53}+0,13*c_{54}+0,13*c_{55}+0,13*c_{56}+0,05*c_{57}+0,05*c_{58}+0,05*c_{59}+0,05*c_{60}+0,05*c_{61}+0,05*c_{62}+0,05*c_{63}+0,05*c_{64}+0,05*c_{65}+0,05*c_{66}+0,05*c_{67}+0,05*c_{68}+0,05*c_{69}+0,05*c_{70}+0,05*c_{71}+0,05*c_{72}+0,05*c_{73}+0,05*c_{74}+0,05*c_{75}+0,05*c_{76}+0,05*c_{77}+0,05*c_{78}+0,05*c_{79}+0,05*c_{80}+0,05*c_{81}+0,05*c_{82}+0,05*c_{83}+0,05*c_{84}+0,05*c_{85}+0,05*c_{86}+0,05*c_{87}+0,05*c_{88}+0,05*c_{89}+0,05*c_{90}+0,05*c_{91}+0,05*c_{92}+0,05*c_{93}+0,05*c_{94}+0,05*c_{95}$

- $Subject\ To \quad 18 * s_1^{d_1} + 13 * s_2^{d_1} + 26 * s_1^{d_2} + 14 * s_2^{d_2} + 27 * s_1^{d_3} + 24 * s_2^{d_3} <= 45$

Ao final da execução da abordagem proposta, o resumo gerado para o grupo de documentos D usado neste exemplo é composto pelas sentenças $s_1^{d_1}$ e $s_1^{d_3}$. Como essas sentenças possuem o mesmo índice, ou seja, são as primeiras sentenças de seus respectivos documentos, elas são ordenadas com base no índice do grupo de sentenças ao qual elas pertencem. A sentença $s_1^{d_1}$ pertence ao grupo g_1, enquanto que a sentença $s_1^{d_3}$ pertence ao grupo g_2. Dessa forma, o resumo é gerado com a sentença $s_1^{d_1}$ em primeiro lugar, seguida pela sentença $s_1^{d_3}$.

> *Hurricane Gilbert swept toward Jamaica yesterday with 100-mile-an-hour winds and officials issued warnings to residents on the southern coasts of the Dominican Republic, Haiti and Cuba. Hurricane Gilbert slammed into Kingston on Monday with torrential rains and 115 mph winds that ripped roofs off homes and buildings, uprooted trees and downed power lines.*

5.2 Experimentos

Vários experimentos foram realizados visando analisar os seguintes aspectos: **(i)** Avaliação da adoção de diferentes formas de representação e métodos de ponderação de conceitos (Seção 5.2.2); **(ii)** Análise do impacto dos limiares de similaridade das sentenças λ e do tamanho mínimo do grupo de sentenças γ adotados nas etapas de agrupamento das sentenças e geração do resumo, respectivamente (Seção 5.2.3); e **(iii)** Comparação dos resultados obtidos com a abordagem proposta em relação a outros sistemas do estado da arte (Seção 5.2.4).

5.2.1 Configurações dos Experimentos

A abordagem proposta é avaliada adotando os corpora das competições do DUC dos anos de 2001 até 2004. Os corpora do DUC são amplamente utilizados para avaliar sistemas de SAT de artigos de notícias escritos em Inglês. Todos os grupos de documentos do DUC possuem um ou mais resumos de referência (com aproximadamente 100 palavras)

criados por especialistas humanos. Algumas estatísticas básicas desses corpora gerados aplicando a ferramenta Stanford CoreNLP são apresentadas na Tabela 21.

Tabela 21 – Estatísticas básicas dos corpora do DUC 2001-2004.

Corpus	#Grupos	#Documentos	#Sentenças	#Palavras
DUC 2001	30	309	11.026	269.990
DUC 2002	59	576	14.370	348.012
DUC 2003	30	298	7.691	197.483
DUC 2004	50	500	13.135	336.073

As medidas de cobertura do ROUGE-1 (R-1) e ROUGE-2 (R-2) (LIN, 2004) são adotadas em todos os experimentos realizados. Essas duas medidas computam a sobreposição de unigramas e bigramas, respectivamente, entre os resumos gerados automaticamente e o conjunto de resumos de referência. Essas medidas foram adotados porque demonstraram ter uma alta correlação com avaliações humanas na literatura (LIN, 2004; OWCZARZAK et al., 2012). A versão ROUGE 1.5.5 foi executada, adotando o seguintes parâmetros: *-m -l 100 -f A*.

Em todos os experimentos realizados, o limiar do tamanho máximo do resumo a ser gerado foi definido como 105 palavras. Esse limiar foi adotado por possibilitar a geração de resumos com tamanhos equivalentes aos produzidos pelos sistemas do estado da arte considerados nos experimentos (aproximadamente 100 palavras). Todas as avaliações foram executadas em um computador com as seguintes especificações: Intel Core i7-4510U com 2,6 GHz, 16 gigabytes de memória RAM, um terabyte de disco rígido e executando o Windows 8.1 64 bits.

É importante ressaltar que os resultados dos experimentos discutidos neste capítulo para o corpus do DUC 2003 diferem em relação aos apresentados no Capítulo 3. Essa diferença ocorre porque mais resumos de referência escritos pelos organizadores da competição do DUC 2003 foram adotados nos experimentos deste capítulo para a avaliação dos resumos automáticos nesse corpus.

5.2.2 Avaliando as Formas de Representação e os Métodos de Ponderação de Conceitos

Este primeiro experimento avalia o desempenho da adoção de cinco formas de representação de conceitos. Além das formas tradicionais usando unigramas ou bigramas, também consideramos as seguintes representações investigadas por Schluter e Søgaard (2015): entidades nomeadas, dependências sintáticas rotuladas e com um rótulo genérico *dep*.

Entidades Nomeadas: A ideia dessa representação é utilizar como conceitos expressões que se referem a nomes de pessoas, lugares, organizações, entre outros. Tais elementos são importantes para sumarização, pois descrevem entidades do mundo real que são mencionadas no documento. Além dessas classes tradicionais, outras categorias classificadas como entidades pela ferramenta Stanford CoreNLP também foram consideradas: Porcentagem, Data, entre outras.

Dependências Sintáticas: Essa representação utiliza as dependências sintáticas entre as palavras como conceitos. Algumas dessas dependências podem ser do tipo sujeito, objeto direto e indireto, complemento, entre outras. Por exemplo, dada a frase "*John walks on the beach.*", as seguintes dependências são extraídas: *root(ROOT, walks)*, *nsubj(walks, John)*, *case(beach, on)*, *det(beach,the)*, *nmod:on(walks, beach)*. Assim como proposto por Schluter e Søgaard (2015), duas formas de representação são derivadas a partir das dependências sintáticas: **(i)** Usando explicitamente o tipo da dependência para rotular a relação (Dep. Sintática R.), por exemplo, *nsubj(walks, John)*; e **(ii)** Adotando um tipo genérico para rotular as dependências (Dep. Sintática G.), por exemplo, *dep(walks, John)*.

O método de ponderação de conceitos proposto e descrito na Seção 5.1.1 e os seguintes métodos individuais são avaliados em conjunto com as cinco representações descritas acima. Vale salientar que todos os escores de relevância são normalizados e variam entre 0 e 1.

Frequência do Conceito: Neste método, o peso de um conceito é dado pelo número de vezes que ele é mencionado na coleção de documentos de entrada, normalizado pelo total de conceitos dos documentos.

Frequência do Conceito - Frequência Inversa do Conceito (FC-FIC): Este método é baseado no tradicional TF-IDF muito usado na área de recuperação de informação. O FC-FIC de um conceito c_i é calculado como demonstrado na Equação 5.4.

$$FC - FIC(c_i) = FC(c_i) \times log\left(\frac{S}{s_{c_i}}\right) \tag{5.4}$$

no qual,

- FC retorna a frequência de um conceito c_i em todos os documentos de entrada;
- S é o total de sentenças na coleção de documentos;
- s_{c_i} é o total de sentenças nas quais c_i é mencionado.

Frequência dos Documentos: Este método computa a importância de um conceito com base na quantidade de documentos que o citam.

Frequência das Sentenças: Neste método, a relevância de um conceito é dado pelo número de frases em que ele aparece.

Método Proposto: É o método de ponderação do conceito proposto neste trabalho e apresentado na Equação 5.1.

Posição das Sentenças: A importância de um conceito é dada com base na média da pontuação da posição da primeira sentença que menciona o conceito em cada documento, conforme apresentado na Equação 5.2.

Os limiares de similaridade λ usado na etapa de agrupamento das sentenças e do tamanho mínimo do grupo de sentenças γ empregado na etapa de geração de resumo são definidos como zero neste experimento, ou seja, nenhuma sentença é removida nessas etapas usando esses limiares. Na seção 5.2.3, o impacto desses limites no desempenho da abordagem proposta é analisado.

São apresentados na Tabela 22 e na Tabela 23 os resultados da avaliação das formas de representação e dos métodos de ponderação de conceito em termos das medidas de cobertura do ROUGE-1 (R-1) e ROUGE-2 (R-2). Cenários de diferenças com significância estatística entre o método com melhor desempenho e os demais métodos avaliados são indicadas usando o símbolo †. Para isso, utilizou-se o teste *Wilcoxon signed-rank* adotando 95% de nível de significância para ambas as medidas do ROUGE.

No total, 48 comparações foram realizadas levando em consideração os quatro corpora adotados, os seis métodos de ponderação de conceitos investigados e as duas medidas de avaliação do R-1 e R-2. Analisando as formas de representação de conceito, adotando bigramas resultou em um melhor desempenho em relação às demais representações em 56,25% das comparações, seguido pela Dep. Sintática G. (20,83%), Dep. Sintática R. (14,58%) e unigramas (8,34%). Além dos melhores resultados com base nas medidas do ROUGE, a utilização de bigramas introduz menos variáveis no modelo de PLI do que outras representações, como o unigramas, e também requer menor esforço computacional durante a etapa de pré-processamento do que adotar dependências sintáticas rotuladas ou não rotuladas. A adoção de Entidades Nomeadas como conceitos apresentou o pior desempenho entre as formas de representação de conceito avaliados. Uma possível razão para esse baixo desempenho é a pouca quantidade de entidades reconhecidas pela ferramenta Stanford CoreNLP.

Analisando os métodos individuais de ponderação de conceitos, o método de posição das sentenças e frequência dos documentos apresentaram, no geral, os dois melhores resultados em ambas as medidas do ROUGE. O método de posição das sentenças obteve melhor desempenho do que todos os outros métodos em dois dos quatro conjuntos de dados (DUC 2002 e DUC 2004) em ambas as medidas do R-1 e R-2, e no DUC 2001 com base na pontuação do R-2. O método de frequência dos documentos apresentou o melhor resultado no DUC 2001 em relação à pontuação R-1. Os métodos de frequência

Tabela 22 – Resultados (%) e desvio padrão (entre parênteses) no DUC 2001-2002. O melhor desempenho global de cada corpus é destacado em negrito, e o grupo de configurações estatisticamente semelhantes, se existir, é indicado pelo símbolo †.

Métodos		**DUC 2001**		**DUC 2002**	
		R-1	R-2	R-1	R-2
Freq. Conceito	Bigrama	32,39 (6,60)	7,96 (3,85)	34,29† (5,06)	7,91† (3,37)
	Dep. Sintática R.	31,85† (6,00)	7,16 (3,88)	34,71† (5,42)	8,16† (3,40)
	Dep. Sintática G.	32,09 (5,81)	7,18 (3,79)	34,53† (5,39)	8,21 (3,25)
	Ent. Nomeada	28,72† (5,28)	4,55† (2,05)	32,35† (5,56)	6,32† (3,72)
	Unigrama	31,12† (6,00)	5,21† (2,80)	33,73† (5,75)	6,83† (3,19)
FC-FIC	Bigrama	31,79† (6,40)	6,81 (3,45)	34,21† (5,02)	7,78† (3,62)
	Dep. Sintática R.	32,89 (6,51)	7,61 (3,86)	34,45† (5,82)	7,83† (3,53)
	Dep. Sintática G.	31,97† (6,07)	7,31 (3,46)	34,46† (5,46)	8,04† (3,49)
	Ent. Nomeada	28,33† (5,46)	4,15† (2,04)	31,54† (6,11)	6,00† (3,74)
	Unigrama	30,84† (5,62)	5,17† (2,98)	32,82† (5,60)	6,52† (2,85)
Freq. Documentos	Bigrama	33,50 (6,13)	8,00 (3,79)	35,00† (5,73)	8,33† (4,15)
	Dep. Sintática R.	33,25 (6,62)	7,30 (3,69)	34,64† (5,15)	8,03† (3,89)
	Dep. Sintática G.	32,69 (6,34)	6,91 (3,52)	34,64† (5,03)	8,14† (3,84)
	Ent. Nomeada	29,62† (5,39)	4,55† (2,89)	30,89† (4,87)	5,52† (3,10)
	Unigrama	31,13† (5,36)	4,99† (2,90)	33,20† (6,04)	6,47† (3,82)
Freq. Sentenças	Bigrama	32,15 (5,76)	7,84 (3,78)	34,10† (5,11)	7,93† (3,26)
	Dep. Sintática R.	32,15 (6,53)	7,38 (4,01)	34,69† (5,35)	8,17† (3,41)
	Dep. Sintática G.	32,16 (5,97)	7,23 (3,87)	34,52† (5,17)	8,30 (3,28)
	Ent. Nomeada	28,62† (5,34)	4,48† (2,08)	32,49† (5,56)	6,39† (3,75)
	Unigrama	30,78† (4,77)	4,48† (1,94)	34,87† (5,89)	7,25† (3,85)
Posição Sentenças	Bigrama	32,54 (5,70)	7,24 (3,79)	**36,20** (5,92)	9,04 (4,24)
	Dep. Sintática R.	32,74 (6,66)	**8,38** (4,68)	35,53 (5,81)	8,64 (4,07)
	Dep. Sintática G.	32,72 (5,93)	8,26 (4,30)	35,31 (5,67)	8,40 (4,17)
	Ent. Nomeada	29,65† (5,55)	4,91† (3,04)	32,06† (6,23)	5,95† (3,64)
	Unigrama	31,47† (6,08)	5,70† (3,54)	34,53† (5,88)	7,46† (3,64)
Método Proposto	Bigrama	**33,92** (6,76)	7,94 (3,89)	35,89 (4,91)	**9,15** (3,95)
	Dep. Sintática R.	32,46 (5,93)	7,20 (3,87)	35,78 (5,58)	8,70 (4,27)
	Dep. Sintática G.	32,30 (5,96)	7,25 (3,76)	35,48 (5,63)	8,59 (4,11)
	Ent. Nomeada	30,76† (4,93)	5,27† (2,49)	33,96† (5,21)	7,24† (3,62)
	Unigrama	32,70 (6,02)	5,39† (3,36)	35,76 (5,68)	8,16† (3,79)

das sentenças e FC-FIC obtiveram a melhor performance no DUC 2003 com base no R-1 e R-2, respectivamente.

O método de frequência dos documentos, em geral, é considerado como o melhor método de ponderação para abordagens baseada em conceitos (GILLICK et al., 2009; BOUDIN; MOUGARD; FAVRE, 2015). Contudo, nos experimentos realizados neste capítulo, o método de posição de sentenças apresentou melhores resultados do que o método de frequência dos documentos. Esses resultados demonstram que a posição e a centralidade desempenham um papel fundamental na identificação de conceitos relevantes para compor o resumo gerado a partir de múltiplos artigos de notícias.

Tabela 23 – Resultados (%) e desvio padrão (entre parênteses) no DUC 2003-2004. O melhor desempenho global de cada corpus é destacado em negrito, e o grupo de configurações estatisticamente semelhantes, se existir, é indicado pelo símbolo †.

Métodos		**DUC 2003**		**DUC 2004**	
		R-1	R-2	R-1	R-2
Freq. Conceito	Bigrama	38,78 (5,77)	9,77† (4,00)	37,58† (4,48)	9,22 (3,14)
	Dep. Sintática R.	38,73 (6,68)	10,18† (4,73)	36,94† (4,81)	8,92† (3,25)
	Dep. Sintática G.	38,79 (7,01)	10,22† (4,88)	36,55† (4,84)	8,73† (3,05)
	Ent. Nomeada	36,08† (5,42)	8,43† (3,69)	33,61† (4,86)	6,38† (2,72)
	Unigrama	37,79† (6,37)	9,15† (4,40)	36,44† (4,39)	7,68† (2,95)
FC-FIC	Bigrama	39,13 (5,74)	9,64† (4,06)	37,59† (4,07)	9,10† (2,97)
	Dep. Sintática R.	38,36† (6,45)	10,03† (4,43)	36,93† (4,42)	8,88† (2,98)
	Dep. Sintática G.	39,02 (6,87)	10,44 (4,91)	37,32† (4,28)	8,95† (3,03)
	Ent. Nomeada	36,21† (5,52)	8,57† (3,76)	33,26† (4,62)	6,09† (2,82)
	Unigrama	37,90† (6,45)	8,58† (4,40)	35,91† (5,20)	7,51† (3,07)
Freq. Documentos	Bigrama	39,22 (5,67)	9,46† (3,77)	38,23 (3,38)	9,43 (2,77)
	Dep. Sintática R.	39,40 (6,52)	10,30 (4,69)	36,79† (4,03)	9,04† (3,04)
	Dep. Sintática G.	39,31 (6,47)	10,28 (4,58)	36,79† (3,92)	8,98† (2,90)
	Ent. Nomeada	36,66† (6,05)	8,79† (4,30)	32,24† (4,67)	5,35† (2,49)
	Unigrama	38,46† (6,17)	9,45† (4,22)	35,39† (5,46)	7,51† (3,49)
Freq. Sentenças	Bigrama	39,55 (5,86)	9,95† (4,22)	37,70† (4,17)	9,12† (2,89)
	Dep. Sintática R.	38,60† (6,64)	10,16† (4,73)	37,17† (4,41)	8,89† (3,24)
	Dep. Sintática G.	38,85 (6,83)	10,21† (4,87)	36,89† (4,58)	8,73† (3,18)
	Ent. Nomeada	36,08† (5,42)	8,43† (3,69)	33,78† (4,78)	6,43† (2,70)
	Unigrama	38,07† (6,11)	8,65† (4,25)	36,09† (4,83)	7,55† (3,00)
Posição Sentenças	Bigrama	39,05 (6,41)	9,79† (4,59)	38,46 (4,12)	9,65 (2,76)
	Dep. Sintática R.	38,72 (6,95)	10,17† (4,77)	37,92 (4,28)	9,23 (3,04)
	Dep. Sintática G.	39,19 (6,79)	10,34 (4,90)	38,25 (4,47)	9,45 (3,05)
	Ent. Nomeada	37,39† (5,84)	9,62† (4,60)	35,45† (4,91)	7,59† (2,98)
	Unigrama	39,36 (6,49)	10,24† (5,12)	36,53† (5,33)	8,06† (3,26)
Método Proposto	Bigrama	39,69 (5,94)	10,80 (4,14)	**38,59** (4,47)	**10,10** (3,20)
	Dep. Sintática R.	40,27 (6,13)	10,88 (4,40)	38,09 (4,79)	9,49 (2,88)
	Dep. Sintática G.	40,30 (6,09)	**11,00** (4,40)	38,45 (4,85)	9,56 (2,93)
	Ent. Nomeada	36,87† (5,42)	9,04† (3,99)	36,19† (4,55)	7,59† (3,00)
	Unigrama	**40,43** (6,37)	10,40 (5,26)	37,09† (5,13)	8,33† (3,21)

O método de ponderação de conceitos proposto apresenta melhores resultado do que as técnicas individuais em quase todos os cenários avaliados nas medidas do R-1 e R-2. As únicas duas exceções aconteceram no DUC 2002 e DUC 2001 com base no R-1 e R-2, respectivamente. Nestes dois casos, o método de posição das sentenças obteve melhor desempenho.

5.2.3 Avaliando o Impacto dos Limiares de Similaridade e Tamanho Mínimo dos Grupos de Sentenças

Este experimento avalia o impacto dos limiares de similaridade das sentenças λ e do tamanho mínimo dos grupos de sentenças γ adotado nas etapas de agrupamento de sentenças e geração do resumo, respectivamente. Esses dois limiares são de suma importância, pois são responsáveis pela remoção das sentenças com base na centralidade de suas informações.

Na Figura 7 são apresentados os resultados, com base na medida do R-1, obtidos pela abordagem proposta nos conjuntos de dados do DUC 2001-2004, variando o valor de λ de 0,0 até 1,0. Os resultados obtidos em todos os conjuntos de dados demonstram que, à medida que o valor λ aumenta, uma diminuição significativa na medida R-1 é observada. Esse comportamento ocorre porque, quanto maior o valor de λ, maior o número de frases removidas durante a etapa de agrupamento. Por exemplo, se $\lambda = 0,5$, então, para que uma frase seja atribuída ao grupo com o qual tem maior similaridade, o escore médio de similaridade deve ser superior a 0,5; caso contrário, a sentença é descartada. Com base nesses resultados, o melhor desempenho foi alcançado definindo $\lambda = 0.1$, uma vez que removeu algumas frases e manteve praticamente o mesmo de desempenho que a configuração padrão $\lambda = 0.0$.[3]

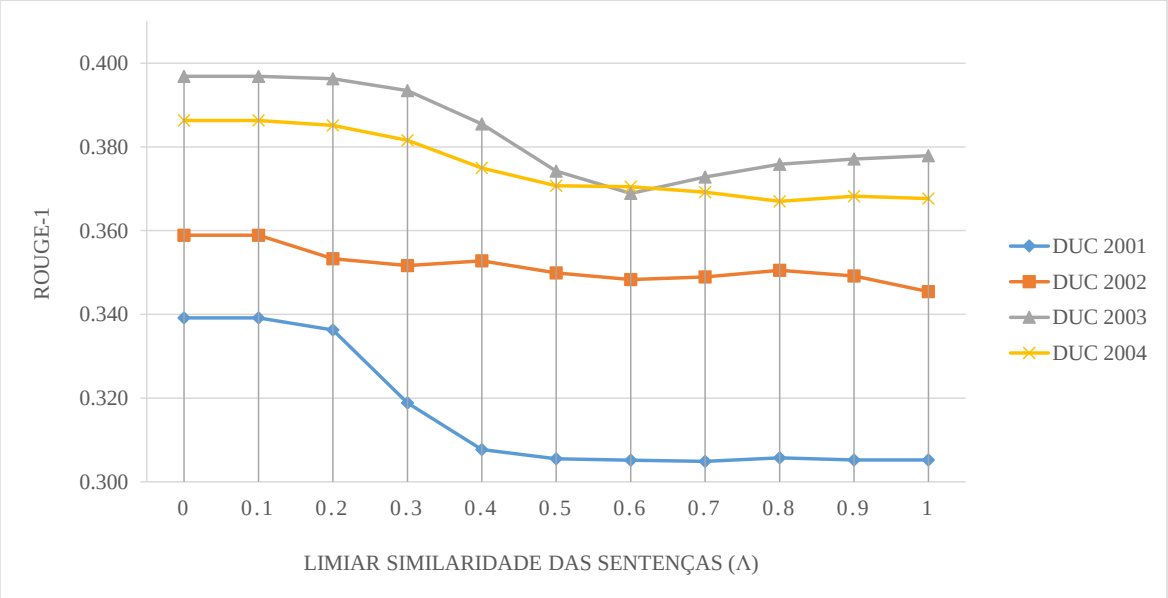


Figura 7 – Impacto do limiar de similaridade mínimo das sentenças λ na medida do R-1.

O limiar γ é usado para remover grupos de sentenças com base no número de membros que o grupo possui. O pressuposto é que quanto mais sentenças um grupo possui, mais importante são as informações que as sentenças desse grupo contêm. Na Figura 8, são apresentados os resultados com base na medida do R-1 obtidos adotando $\lambda = 0,1$ e

[3] Observou-se uma diferença apenas no DUC 2004, em relação às medidas de cobertura do R-1 e R-2.

variando os valores de γ de 0,0 até 1,0. Esses valores indicam a porcentagem do tamanho mínimo que um grupo deve ter para não ser removido. Por exemplo, com $\gamma = 0,3$ e dada uma coleção com 10 documentos de entrada, o grupo de sentenças deve ter 3 ou mais membros para não ser removido. Quanto maior o valor de γ, maior é a quantidade de grupos e sentenças removidas.

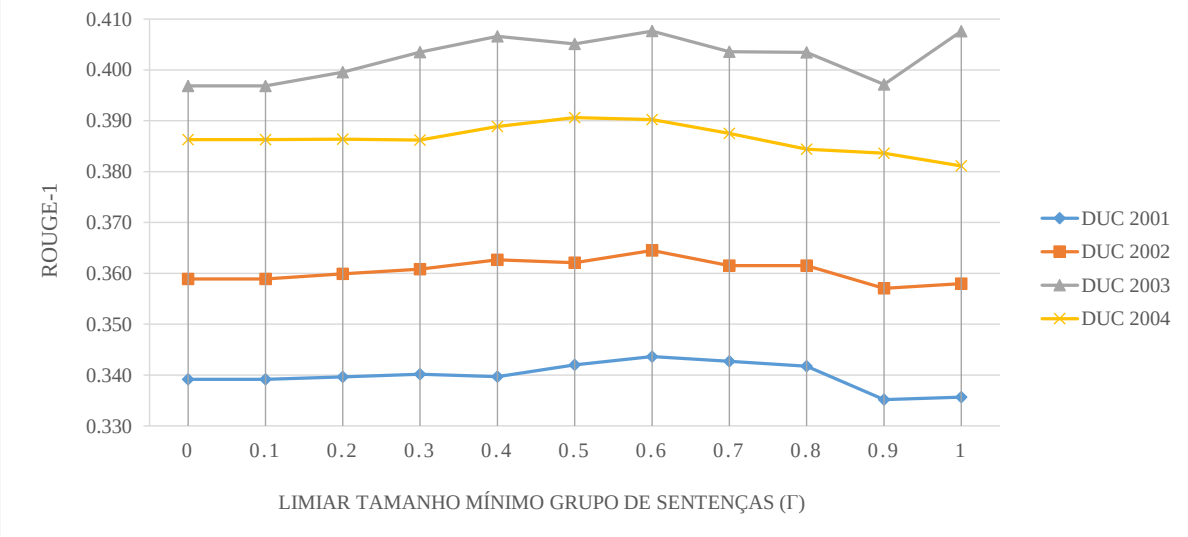


Figura 8 – Impacto do limiar do tamanho mínimo dos grupos de sentenças γ na medida do R-1.

Analisando os resultados obtidos, observa-se que o desempenho da abordagem proposta melhora em termos do R-1 conforme o valor de γ aumenta, até certo valor. Um equilíbrio adequado entre o número total de grupos removidos e os melhores resultados em relação à medida do R-1, considerando o desempenho em todos os conjuntos de dados, foi encontrado com $\gamma = 0,6$.

O número de variáveis e restrições influenciam diretamente a escalabilidade do modelo PLI. McDonald (2007) demonstrou que aplicar PLI para seleção de sentenças na tarefa de sumarização tem uma baixa escalabilidade à medida que o número de variáveis e restrições aumenta. A Figura 9 apresenta os resultados obtidos com a análise do impacto que o limiar γ tem no tempo de execução (em segundos) que o modelo de PLI leva para encontrar uma solução para o processo de seleção de sentenças.

Os resultados demonstram que, em geral, o tempo de execução necessário para que uma solução seja encontrada pelo algoritmo de PLI diminui à medida que o valor de γ aumenta. Vale ressaltar que somente foi computado o tempo gasto pelo modelo de PLI encontrar uma solução; outras etapas da abordagem foram desconsideradas porque não são influenciadas pelo processo de remoção das sentenças. Em alguns casos, observou-se um aumento no tempo de execução, embora o número de variáveis no modelo tenha diminuído. Acreditamos que tal comportamento aconteceu devido à outras questões que também influenciam diretamente no tempo de execução, como a distribuição dos pesos dos conceitos na função objetivo e as restrições adotadas. Definir o limiar do tamanho

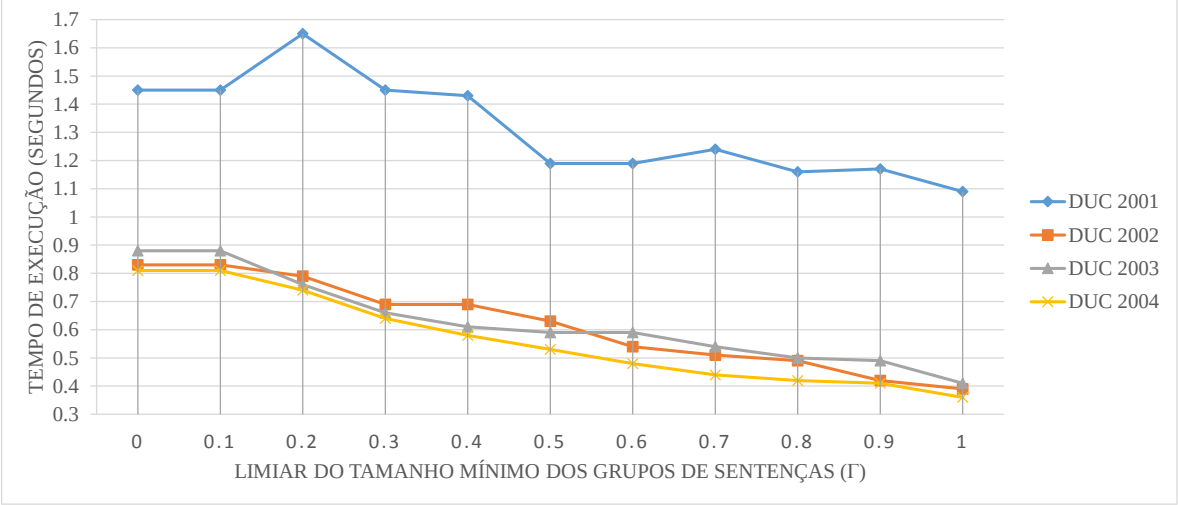


Figura 9 – Impacto do limiar do tamanho mínimo do grupo de sentenças γ no tempo de execução do modelo de PLI.

mínimo dos grupos de sentenças como $\gamma = 0,6$ produziu melhores resultados em relação às medidas do ROUGE e também diminuiu o tempo de execução do algoritmo de PLI em comparação com a configuração padrão adotando $\gamma = 0,0$.

Em resumo, é possível observar nos resultados obtidos que existe um padrão no desempenho com base nas medidas de cobertura do ROUGE, com o método proposto alcançando o melhor desempenho definindo $\lambda = 0,1$ e $\gamma = 0,6$ em todos os corpora avaliados. Tal configuração conduz o método proposto a: **(i)** melhores resultados em relação às medidas do R-1 e R-2 em comparação com a configuração padrão ($\lambda = 0,0$ e $\gamma = 0,0$); e **(ii)** uma menor quantidade de sentenças e conceitos considerados para a construção do modelo de PLI, diminuindo o tempo requerido pelo algoritmo para selecionar as sentenças para compor o resumo.

5.2.4 Comparação com outras abordagens

Esta seção apresenta a comparação do desempenho da abordagem proposta, adotando ($\lambda = 0,1$ e $\gamma = 0,6$) identificado na Seção 5.2.3, em relações aos seguintes trabalhos relacionados: **(i)** os sistemas participantes com o melhor desempenho nas competições DUC 2001-2004 encontrados nos experimentos realizados aqui, e **(ii)** os seguintes sistemas do estado da arte: ICSISumm (GILLICK et al., 2009), Greedy-KL (HAGHIGHI; VANDERWENDE, 2009), LLRSum (CONROY; SCHLESINGER; O'LEARY, 2006), ProbSum (NENKOVA; VANDERWENDE; MCKEOWN, 2006) e Sume (BOUDIN; MOUGARD; FAVRE, 2015). Os resumos do sistema Sume[4] foram gerados usando as configurações padrões disponibilizadas pelos autores, enquanto que, para os outros sistemas, os resumos adotados foram gerados e disponibilizados por Hong, Marcus e Nenkova (2015).

[4] https://github.com/boudinfl/sume

Como pode ser visto na Tabela 24, a abordagem proposta apresentou o melhor desempenho em ambas as medidas do R-1 e R-2 nos corpora do DUC 2001, 2003 e 2004. Enquanto que o sistema ICSISumm obteve a melhor performance no DUC 2002.

Tabela 24 – Resultados (%) e desvio padrão (entre parênteses) das comparações entre a abordagem proposta e outros sistemas em termos das medidas do R-1 e R-2. O melhor desempenho global de cada corpus é destacado em negrito, e o grupo de sistemas estatisticamente semelhantes, se existir, é indicado pelo símbolo †.

Sistemas	**DUC 2001**		**DUC 2002**	
	R-1	R-2	R-1	R-2
Greedy-KL	32,84† (6,43)	6,70 (3,64)	35,79 (5,74)	7,49 (3,61)
ICSISumm	33,88† (6,95)	7,75† (4,30)	**37,34** (5,05)	**9,53** (3,83)
LLRSum	32,00 (5,88)	6,76 (3,25)	32,84 (5,55)	6,75 (3,72)
Sistema DUC	31,69 (6,43)	6,30 (3,76)	35,21 (5,30)	7,66 (3,30)
ProbSum	29,73 (5,41)	5,16 (2,64)	32,57 (4,74)	7,06 (3,63)
Proposta	**34,36** (7,03)	**8,46** (4,23)	36,45† (5,66)	9,27† (3,82)
Sume	33,37† (7,14)	7,73 (4,29)	34,30 (5,26)	8,15 (3,92)
Sistemas	**DUC 2003**		**DUC 2004**	
	R-1	R-2	R-1	R-2
Greedy-KL	40,35† (5,76)	9,20 (3,99)	38,27† (4,73)	8,96 (3,09)
ICSISumm	40,07† (4,88)	10,95 (4,00)	38,42† (4,14)	9,80† (3,17)
LLRSum	36,94 (6,05)	8,87 (3,21)	35,90 (5,01)	8,06 (3,12)
Sistema DUC	38,44 (5,25)	9,11 (3,95)	37,69 (4,08)	8,98 (3,08)
ProbSum	37,60 (7,11)	9,28 (4,05)	35,37 (4,41)	8,18 (3,00)
Proposta	**40,76** (6,10)	**11,22** (4,23)	**39.02** (4.40)	**10.07** (3.27)
Sume	39,36† (5,57)	9,81 (3,95)	37,29 (4,24)	8,83 (2,71)

Analisando apenas os sistemas puramente extrativos, ou seja, excluindo o sistema ICSISumm, a abordagem proposta alcançou melhores resultados do que todos os outros sistemas em todos os cenários e, na maioria deles, com melhorias significativas considerando 95% de nível de confiança. Em particular, esses ganhos foram obtidos na medida do ROUGE-2, demonstrando que o método de ponderação de conceitos proposto melhorou a cobertura dos bigramas relevantes nos resumos gerados. Mais detalhes dos p-valores obtidos com a execução do teste de *Wilcoxon signed rank* são apresentados na Tabela 25, comparando as medidas de cobertura do R-1 dos sistemas avaliados. Os p-valores indicando uma diferença estatisticamente significante ao nível de confiança de 95% são destacados em negrito. Um sinal de maior (>) antes do p-valor indica que o sistema na linha obteve maior média no R-1 do que o sistema na coluna, enquanto que um sinal negativo (<) indica a relação oposta, que o sistema na coluna possui maior R-1 média do que o listado na linha.

Comparando os resultados da abordagem proposta com o sistema ICSISumm, seu desempenho pode ser considerado estatisticamente equivalente nos quatro conjuntos de dados em relação a ambas as medidas do ROUGE. Somente no DUC 2002, o sistema

Tabela 25 – P-valores obtidos aplicando o teste de *Wilcoxon signed rank*.

DUC 2001						
Sistemas	ICSISumm	LLRSum	Sistema DUC	ProbSum	Proposta	Sume
Greedy-KL	(<) 0,149	(>) 0,593	(>) 0,290	(>) **0,002**	(<) 0,092	(<) 0,322
ICSISumm	-	(>) 0,058	(>) **0,012**	(>) **1,05e-04**	(<) 0,895	(>) 0,546
LLRSum	-	-	(>) 0,636	(>) **0,002**	(<) **0,010**	(<) 0,190
Sistema DUC	-	-	-	(>) **0,015**	(<) **0,007**	(<) 0,166
ProbSum	-	-	-	-	(<) **5,74e-05**	(<) **9,58e-04**
Proposta	-	-	-	-	-	(>) 0,106
DUC 2002						
Sistemas	ICSISumm	LLRSum	Sistema DUC	ProbSum	Proposta	Sume
Greedy-KL	(<) **0,011**	(>) **1,50e-04**	(>) 0,425	(>) **4,48e-04**	(<) 0,408	(>) 0,054
ICSISumm	-	(>) **1,19e-06**	(>) **0,001**	(>) **4,78e-08**	(>) 0,155	(>) **1,44e-06**
LLRSum	-	-	(<) **0,002**	(>) 0,958	(<) **3,17e-06**	(<) **0,008**
Sistema DUC	-	-	-	(>) **0,002**	(<) 0,196	(>) 0,239
ProbSum	-	-	-	-	(<) **4,22e-06**	(<) 0,050
Proposta	-	-	-	-	-	(>) **1,66e-04**
DUC 2003						
Sistemas	ICSISumm	LLRSum	Sistema DUC	ProbSum	Proposta	Sume
Greedy-KL	(>) 0,715	(>) **7,92e-04**	(>) **0,022**	(>) **0,001**	(<) 0,641	(>) 0,309
ICSISumm	-	(>) **0,001**	(>) 0,088	(>) 0,062	(<) 0,364	(>) 0,561
LLRSum	-	-	(<) 0,099	(<) 0,424	(<) **0,0001**	(<) **0,003**
Sistema DUC	-	-	-	(>) 0,490	(<) **0,004**	(<) 0,416
ProbSum	-	-	-	-	(<) **0,008**	(<) 0,088
Proposta	-	-	-	-	-	(>) 0,161
DUC 2004						
Sistemas	ICSISumm	LLRSum	Sistema DUC	ProbSum	Proposta	Sume
Greedy-KL	(<) 0,528	(>) **5,87e-05**	(>) 0,291	(>) **1,69e-04**	(<) 0,085	(>) 0,141
ICSISumm	-	(>) **4,10e-05**	(>) 0,114	(>) **5,64e-05**	(<) 0,202	(>) **0,017**
LLRSum	-	-	(<) **0,015**	(>) 0,301	(<) **1,12e-06**	(<) **0,005**
Sistema DUC	-	-	-	(>) **0,003**	(<) **0,015**	(>) 0,779
ProbSum	-	-	-	-	(<) **9,92e-06**	(<) **0,001**
Proposta	-	-	-	-	-	(>) **0,003**

ICSISumm obteve melhores resultados do que a abordagem proposta. Estes resultados são encorajadores, uma vez que a abordagem proposta usa somente as sentenças originais do documento na geração dos resumos, diferentemente do ICSISumm, que aplica técnicas de compressão de sentenças com base na árvore sintática da frase.

Na Tabela 26 são apresentados exemplos de resumos gerados para a coleção de documentos *d061* do corpus do DUC 2002 usando a abordagem proposta, o sistema ICSISumm e um dos resumos de referência escrito pelo especialista humano. Ambos os resumos gerados automaticamente para este grupo apresentam um nível razoável de informatividade com base na medida do ROUGE-1. Contudo, ainda há uma diferença notável no nível de qualidade, tanto em relação à coesão quanto à informatividade, comparado com o resumo produzido pelo perito humano. Esse fato demonstra que há margem para melhorias, especialmente seguindo a evolução das abordagens extrativas para abstrativas.

Tabela 26 – Exemplo de resumos gerados para a coleção *d061* no corpus do DUC 2002.

Sistemas	**Resumos**
ICSISumm	Hurricane Gilbert slammed into Kingston with torrential rains and 115 mph winds that ripped roofs off homes and buildings, uprooted trees and downed power lines. A National Weather Service report said the hurricane was moving west at 17 mph with maximum sustained winds of 115 mph. Gilbert reached Jamaica after skirting southern Puerto Rico, Haiti and the Dominican Republic. “It’s certainly one of the larger systems we’ve seen in the Caribbean for a long time,"said Hal Gerrish, forecaster at the National Hurricane Center. Hurricane warnings were issued for the south coast of Haiti and Cuba by their respective governments.
Proposed Approach	Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas. The storm ripped the roofs off houses and caused coastal flooding in Puerto Rico. It’s certainly one of the larger systems we’ve seen in the Caribbean for a long time,"said Hal Gerrish, forecaster at the National Hurricane Center in Coral Gables, Fla. A National Weather Service report said the hurricane was moving west at 17 mph with maximum sustained winds of 115 mph. Hurricane warnings were posted for the Cayman Islands, Cuba and Haiti.
Human-made	Tropical Storm Gilbert strengthened into a hurricane on Saturday night, September 10th in the eastern Caribbean. It tracked westerly at about 15mph while building in intensity. After skirting southern Puerto Rico, Haiti and the Dominican Republic, it hit Jamaica with high winds and torrential rains, destroying 100,000 of the countries 500,000 homes and taking nineteen lives. It then passed over the Cayman Islands before slamming into the Yucatan Peninsula causing heavy damage in the Cancun and Cozumel regions. Gilbert is the most intense hurricane ever recorded with a record low barometric pressure of 26.31 inches and sustained winds of 179mph, gusting to 218mph.

5.3 Considerações Finais do Capítulo

Este capítulo apresentou uma abordagem baseada em conceitos utilizando PLI para SAT multidocumento de artigos de notícias que explora a integração dos métodos de centralidade e posição para filtrar frases menos relevantes e ponderar a importância dos conceitos extraídos. A abordagem proposta adota uma estratégia baseada em centralidade para executar o processo de agrupamento das sentenças com o objetivo de remover as frases com um baixo grau de centralidade e auxiliar no processo de ordenação das sentenças no resumo gerado.

Os resultados experimentais demonstram a eficácia da abordagem proposta nos quatro corpora das competições do DUC 2001-2004, com base nas medidas de cobertura do

ROUGE-1 e ROUGE-2. A integração dos métodos de centralidade e posição demonstrou ser eficiente, fazendo com que a abordagem proposta obtivesse resultados competitivos com diversos sistemas do estado da arte. A estratégia de filtragem das sentenças baseada no percentual de documentos de entrada provou ter um impacto positivo, aumentando a informatividade dos resumos gerados e também diminuindo o tempo de execução do modelo de PLI.

Apesar dos resultados encorajadores obtidos pela abordagem proposta, ela ainda é estática, ou seja, uma mesma configuração de parâmetros e métodos é usada para toda coleção de documentos de entrada. Conforme observado por Hong, Marcus e Nenkova (2015) essa é uma importante limitação, já que nenhum método de sumarização consegue produzir resumos informativos para todos os grupos de documentos de entrada, mesmo quando eles pertencem ao mesmo domínio. Essa limitação pode ser comprovada observando o razoável valor do desvio padrão no desempenho da abordagem proposta, o que demonstra que ela não conseguiu gerar resumos informativos para todas as coleções de documentos de entrada. Vislumbrando mitigar essa limitação, no próximo capítulo apresentamos uma abordagem baseada em conceitos para a sumarização monodocumento e multidocumento usando PLI e regressão.

6 UMA ABORDAGEM BASEADA EM CONCEITOS VIA PLI E REGRESSÃO

As abordagens propostas no Capítulo 4 e Capítulo 5 apresentaram resultados competitivos com diversos sistemas do estado da arte nas tarefas de sumarização monodocumento e multidocumento, respectivamente. Contudo, ambas as soluções propostas ainda são estáticas, ou seja, uma configuração pré-definida (método de ponderação, forma de representação) é adotada para sintetizar todo documento ou coleção de documentos de entrada. A revisão bibliográfica apresentada no Capítulo 2 evidenciou que a maioria dos trabalhos identificados em ambas as tarefas também possuem essa característica. Análises apresentadas na literatura (HONG; MARCUS; NENKOVA, 2015; WAN et al., 2015) evidenciam que a aplicação de uma única abordagem de sumarização extrativa para todo(s) documento(s) de entrada é uma significante limitação. Isso acontece porque um único método de sumarização extrativo não consegue manter uma alta performance para qualquer documento a ser sumarizado, mesmo quando eles pertencem ao mesmo domínio. Tal comportamento é esperado, dadas a subjetividade e a complexidade da tarefa de criação de resumos.

Alguns trabalhos, como Ferreira et al. (2013), Ferreira et al. (2014), investigaram e concluíram que as técnicas de sumarização apresentam desempenhos distintos com base no tipo de documento a ser sumarizado. Dessa forma, os autores propõem a utilização de técnicas diferentes, dependendo do tipo do documento, (por exemplo, artigos científicos, blogs, artigos de notícias, dentre outros). Contudo, analisando os resultados obtidos neste trabalho e em outros trabalhos (HONG; MARCUS; NENKOVA, 2015; WAN et al., 2015), essa conclusão pode ser complementada, já que a variabilidade no desempenho dos métodos de sumarização é alta mesmo em documentos de um único domínio.

Na tentativa de suprir essa limitação, algumas abordagens propõem a estratégia de primeiro gerar vários resumos para cada documento ou coleção de entrada, e em uma segunda etapa, realizar o processo de combinação (PEI et al., 2012; WANG; LI, 2012; FERREIRA et al., 2014; MEENA; DEOLIA; GOPALANI, 2015), ordenamento (WAN et al., 2015) ou estimação da informatividade (HONG; MARCUS; NENKOVA, 2015) desses resumos candidatos. Abordagens centradas na combinação de diversos resumos (PEI et al., 2012; WANG; LI, 2012; FERREIRA et al., 2014; MEENA; DEOLIA; GOPALANI, 2015) têm apresentado resultados melhores do que adotar os resumos individualmente. No entanto, não existe nenhuma garantia de que combinar dois ou mais resumos resultará em um terceiro resumo mais informativo. Hong, Marcus e Nenkova (2015) demonstraram que adotar algoritmos de regressão para estimar a informatividade de um resumo apresentou melhores resultados do que adotar algoritmos de ranqueamento para a sumarização multidocumento.

A estratégia de estimar a informatividade de um resumo aplicando algoritmos de

regressão permite quantificar a sua informatividade como um todo por meio de um valor contínuo. Dessa forma, é possível realizar uma melhor análise de cada resumo do que ranqueá-los ou combiná-los. No melhor do conhecimento do autor desta tese, o trabalho mais similar ao proposto neste capítulo é o sistema SumCombine introduzido por Hong, Marcus e Nenkova (2015) para a sumarização multidocumento. Enquanto isso, nenhum trabalho que tenha adotado algoritmos de regressão para estimar a informatividade de um resumo na tarefa de sumarização monodocumento foi encontrado.

Hong, Marcus e Nenkova (2015) propuseram o sistema SumCombine para sumarização multidocumento que gera diferentes resumos candidatos combinando em nível de sentença os resumos produzidos por quatro sistemas do estado da arte. Em uma segunda etapa, o algoritmo de máquina de vetores de suporte para regressão é aplicado para estimar a informatividade de cada resumo candidato. Por fim, o resumo estimado como mais informativo é selecionado. Duas limitações importantes podem ser ressaltadas na abordagem proposta, sendo elas:

1. O processo de geração dos resumos candidatos produz uma quantidade muito grande de candidatos devido às diversas combinações em nível de sentenças realizadas. Isso resulta em um elevado custo computacional para gerar e analisar cada candidato. Além disso, por ser uma combinação de forma desordenada, diversos resumos com baixa informatividade são gerados.

2. O modelo de regressão construído adota medidas de divergência e similaridade que somente levam em consideração a distribuição de probabilidades dos n-gramas (unigramas e bigramas) do resumo e do conjunto de documentos de entrada. Nessas medidas, a ausência ou presença de n-gramas com maior probabilidade apresenta maior impacto, ou seja, a probabilidade de ocorrência de um n-grama funciona como um quantificador da sua importância. Outros aspectos importantes para a SAT, como posição e centralidade, não são considerados.

A abordagem proposta neste capítulo visa suprir as limitações supracitadas do método proposto por Hong, Marcus e Nenkova (2015), além de estendê-la para a sumarização monodocumento. Neste contexto, este capítulo apresenta uma abordagem baseada em conceitos utilizando programação linear inteira e regressão para as tarefas de sumarização monodocumento e multidocumento de artigos de notícias. A arquitetura da solução proposta é composta por duas etapas principais: *Geração dos resumos candidatos* e *Seleção do resumo mais informativo*.

O processo de geração dos resumos candidatos é realizado adotando as abordagens propostas no Capítulo 4 e no Capítulo 5 para a sumarização monodocumento e multidocumento, respectivamente. Para isso, diferentes métodos de ponderação e formas de representação de conceitos são explorados, visando gerar uma grande variedade de resumos de candidatos.

Em seguida, um algoritmo de regressão é aplicado para estimar a informatividade de cada resumo candidato usando a tradicional medida de cobertura do ROUGE-1 (LIN, 2004) como atributo alvo. O modelo de regressão proposto é treinado com várias características identificadas na literatura e os novos atributos propostos neste trabalho. Os aspectos de posição, centralidade e frequência são usados em conjunto com diversas medidas de similaridade e divergência/distância para estimar a informatividade dos resumos sob diferentes perspectivas.

Este capítulo reexamina a hipótese de que é possível estimar a informatividade de um resumo adotando um conjunto de características extraídas do próprio resumo e com base em medidas que comparam a similaridade e a divergência entre o(s) documento(s) de entrada e o resumo. Hong, Marcus e Nenkova (2015), Wan et al. (2015) dão suporte a esta hipótese diante dos resultados encorajadores encontrados adotando algumas das características investigadas aqui para a tarefa de sumarização multidocumento.

Em uma linha de pesquisa diferente, voltada para a avaliação de sistemas de SAT, Louis e Nenkova (2009), Saggion et al. (2010) investigaram a estratégia de avaliar sistemas de sumarização sem utilizar resumos de referência. Para isso, os autores exploraram diversas características extraídas dos resumos gerados e das relações de similaridade e divergência com os documentos originais. Ambos os trabalhos obtiveram resultados encorajadores, demonstrando que existia uma boa correlação das características investigadas com as medidas de avaliação do PYRAMID e do ROUGE.

As principais contribuições deste capítulo são:

- Uma abordagem baseada em conceitos utilizando programação linear inteira e regressão para as tarefas de sumarização monodocumento e multidocumento. O método proposto adota uma abordagem baseada em conceitos com PLI para gerar diversos resumos candidatos e, em seguida, um algoritmo de regressão é aplicado para selecionar o resumo mais informativo.

- Um conjunto de características projetadas com base em indicadores individuais e combinados de importância de conteúdo, como posição, frequência e centralidade. As características propostas obtiveram correlações mais fortes com a medida de cobertura do ROUGE-1 do que outros atributos adotados anteriormente na literatura baseados em probabilidade.

- Uma eficaz estratégia para a geração de resumos candidatos que explora diferentes métodos de ponderação de conceitos, formas de representação e outras configurações, integradas com as abordagens propostas no Capítulo 4 e Capítulo 5. Tal estratégia permite a geração de uma grande diversidade de resumos candidatos informativos, conduzindo a elevados limites superiores com base nas medidas de cobertura do ROUGE-1 e ROUGE-2.

O restante deste trabalho está organizado da seguinte forma. A Seção 6.1 descreve todas as etapas da abordagem proposta. Na Seção 6.2 são apresentados e discutidos os experimentos realizados para avaliar diferentes aspectos da abordagem proposta. Finalmente, a Seção 6.3 apresenta as considerações finais deste capítulo.

6.1 Abordagem Proposta

A metodologia de sumarização desenvolvida neste capítulo propõe a integração da geração de diversos resumos candidatos, aplicando uma abordagem baseada em conceitos usando PLI, com uma análise posterior para estimar e selecionar o resumo mais informativo. A Figura 10 apresenta uma visão geral da abordagem proposta, que consiste nas seguintes etapas: Pré-Processamento, Geração dos resumos candidatos e Seleção do resumo mais informativo.

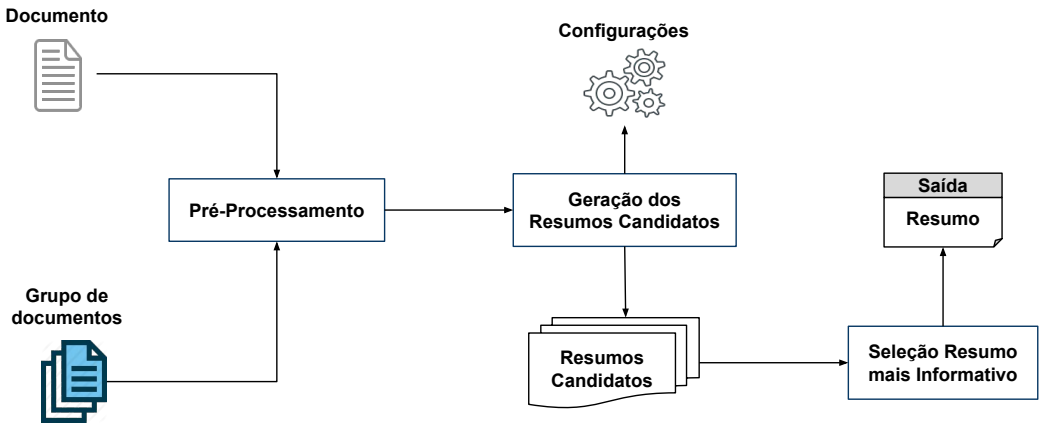


Figura 10 – Visão geral da abordagem proposta via PLI e regressão.

Pré-processamento: O documento ou a coleção de documentos textuais de entrada são pré-processados, aplicando a ferramenta *Stanford CoreNLP* (MANNING et al., 2014). As seguintes tarefas de processamento de linguagem natural são executadas: tokenização, segmentação das sentenças, etiquetagem gramatical das palavras, lematização, análise sintática e reconhecimento de entidades nomeadas, e resolução de correferências.

Geração dos resumos candidatos: Esta etapa é responsável pela geração do conjunto de resumos candidatos. Para tal, as abordagens propostas no Capítulo 4 e no Capítulo 5 são adotadas, dependendo se a tarefa de sumarização for monodocumento ou multidocumento, respectivamente. Diversas formas de representação e métodos de ponderação de conceitos são adotados, visando gerar uma grande variedade de resumos candidatos.

Seleção do resumo mais informativo: Nesta etapa, cada resumo candidato gerado anteriormente é analisado, visando estimar o quão informativo ele é. Um modelo de regressão é treinado com diversas características que capturam a cobertura de informações

relevantes presentes no resumo. Por fim, o resumo estimado como mais informativo é selecionado como o mais adequado para representar o documento ou a coleção de documentos de entrada.

Uma descrição detalhada das etapas de *Geração dos resumos candidatos* e *Seleção do resumo mais informativo* é apresentada nas próximas seções.

6.1.1 Geração dos resumos candidatos

Esta etapa é responsável pela geração de um conjunto de resumos candidatos a partir de um único documento (monodocumento) ou de uma coleção de documentos (multidocumento) de entrada. Para isso, as abordagens baseadas em conceitos usando PLI apresentadas no Capítulo 4 e no Capítulo 5 são utilizadas, dependendo da tarefa de sumarização a ser executada.

Dois parâmetros fundamentais que precisam ser definidos em uma abordagem baseada em conceitos usando PLI são: **(i)** uma forma de representação para a noção de um conceito (c_i); e **(ii)** um método para ponderar a sua relevância (w_i). Esta etapa explora as diferentes formas de representação e métodos de ponderação investigadas até então neste trabalho, visando gerar um conjunto diversificado de resumos candidatos. Além disso, outros parâmetros específicos de cada uma das abordagens também são adotados para variar o processo de sumarização.

O Algoritmo 3 resume a metodologia de geração dos resumos candidatos executada nesta etapa. O primeiro ponto a ser definido é qual a tarefa de sumarização que será realizada. Caso o conjunto de documentos D de entrada possua apenas um único documento $|D| = 1$, a abordagem para sumarização monodocumento descrita no Capítulo 4 é adotada, caso contrário, $|D| > 1$, a abordagem multidocumento apresentada no Capítulo 5 é utilizada. Posteriormente, a coleção de configurações CF é adotada em conjunto com a abordagem de sumarização selecionada. Cada configuração $conf_i \in CF$ contém a definição de qual forma de representação e método de ponderação de conceitos serão utilizados. Além disso, outros parâmetros específicos dependendo da tarefa de sumarização (monodocumento ou multidocumento) também são definidos. Por exemplo, considerar ou não a pontuação do grafo de entidades na sumarização monodocumento, ou os valores dos limiares de similaridade mínima entre as sentenças λ e do tamanho mínimo do grupo de sentenças γ para a sumarização multidocumento. Cada configuração considerada produz um resumo e esse é inserido na lista de resumos candidatos.

É essencial para a abordagem proposta que essa etapa seja capaz de gerar um conjunto de resumos candidatos informativos e com uma grande diversidade. Para isso, as seguintes formas de representação de conceitos são adotadas: unigramas, bigramas, entidades nomeadas e dependências sintáticas rotuladas e genéricas. Tais formas de representação apresentaram bons resultados nos experimentos realizados no Capítulo 4, Capítulo 5 e no Apêndice A, em conjunto com as abordagens baseadas em conceitos propostas. Além

Algoritmo 3: Etapa de geração dos resumos candidatos.

Entrada: O conjunto de documentos D.
Entrada: O conjunto de configurações CF.
Saída: O conjunto de resumos candidatos $resumosCandidatos$.

```
1  Início
2      sumarizador = null
3      se |D| == 1 então
4          sumarizador = sumarizadorMonoDoc
5      senão
6          sumarizador = sumarizadorMultiDoc
7      resumosCandidatos = {}
8      para todo conf_i ∈ CF faça
9          sumarizador.configuracao = conf_i
10         resumo = sumarizador.sumarizar(D)
11         resumosCandidatos.add(resumo)
```

disso, observou-se uma alta diversidade nos resumos gerados adotando essas formas de representação.

Os resultados apresentados no Capítulo 3 demonstraram a eficiência de diversos métodos de pontuação de sentenças, em ambas as tarefas de sumarização. Além disso, observou-se que esses métodos geram, em muitos casos, resumos distintos dos produzidos utilizando as cinco formas de representação de conceitos mencionadas anteriormente. Por isso, visando diversificar ainda mais os resumos candidatos gerados, sentenças também são adotadas como conceitos.

Para exemplificar, as seguintes formas para representação da noção de conceitos são usadas nesta etapa.

- **Sentenças**: *Hurricane Gilbert slammed into Kingston on Monday*
- **Unigramas**: *Hurricane - Gilbert - slammed - into - Kingston*
- **Bigramas**: *Hurricane Gilbert - Gilbert slammed - slammed into*
- **Entidades Nomeadas**: *Hurricane Gilbert - Kingston - Monday*
- **Dependências Sintáticas Rotuladas**: *comp*(Gilbert, Hurricane) - *nsubj*(slammed, Gilbert)
- **Dependências Sintáticas Genérica**: *dep*(Gilbert, Hurricane) - *dep*(slammed, Gilbert)

O segundo ponto fundamental em uma abordagem baseada em conceitos é a definição do método de ponderação aplicado para mensurar a relevância w_i de um conceito c_i. Diversas técnicas individuais foram investigadas até o momento neste trabalho. Esses

métodos incluem frequência do conceito, frequência das sentenças, posição das sentenças, frequência dos documentos, métodos baseados em grafos, como, PageRank, TextRank, HITS, entre outros. Além das técnicas individuais, também foram avaliados os métodos combinados propostos para sumarização monodocumento (ver Equação 4.4) e multidocumento (ver Equação 5.2). Todos esses métodos são usados em conjunto com as cinco formas de representação mencionadas anteriormente[1].

A maioria dos métodos de ponderação mencionados anteriormente foram propostos e avaliados usando fragmentos textuais que compõem uma sentença, por exemplo, unigramas e bigramas, como conceitos. Dessa forma, em geral, eles não são adequados para mensurar a relevância de uma sentença completa. Por isso, para ponderar a importância das sentenças como conceitos, os métodos de pontuação de sentenças individuais investigados no Capítulo 3 são adotados.

Abordagens centradas na estratégia de maximizar a cobertura de conceitos relevantes exploram somente a melhor solução obtida pelo modelo de PLI (GILLICK et al., 2009; BOUDIN; MOUGARD; FAVRE, 2015; CAO et al., 2015). A solução ótima selecionada representa um resumo gerado usando o conjunto de sentenças que maximiza a função objetivo e atende todas as restrições do modelo. Esta etapa também explora a execução do modelo de PLI várias vezes, visando obter as **N** soluções distintas em ordem decrescente com base no valor da função objetivo. Para isso, a restrição apresentada na Equação 6.1 é inserida no modelo de PLI das abordagens monodocumento (ver Equação 4.1) e multidocumento (ver Equação 5.3).

$$\sum_{Sol_k \in SOL} \sum_{s_m \in Sol_k} s_m \leq |Sol_k| - 1 \tag{6.1}$$

No qual,

- SOL é o conjunto de soluções obtidas pelo modelo de PLI;
- Sol_k é uma solução específica composta por uma ou mais sentenças;
- $|Sol_k|$ é o total de sentenças da solução Sol_k;
- s_m é uma sentença pertencente a solução Sol_k;

Para ilustrar o funcionamento da restrição acima, imagine que a solução $Sol_1 = \{s_1, s_2, s_3, s_4\}$, composta por quatro sentenças ($|Sol_1| = 4$), foi produzida pelo modelo de PLI. Com isso, em uma segunda execução, a restrição $s_1 + s_2 + s_3 + s_4 \leq 3$ é inserida no modelo. Tal restrição garante que a nova solução gerada possua pelo menos uma sentença diferente da solução Sol_1, ou seja, o novo resumo gerado não pode possuir exatamente as mesmas sentenças do resumo produzido anteriormente. Adotando essa restrição, é possível gerar resumos distintos usando uma mesma configuração.

[1] É importante ressaltar que os métodos baseados em grafos, como PageRank, não foram adotados para a sumarização multidocumento por conta do alto custo computacional observado durante a geração do grafo de conceitos.

O Quadro 12 resume todas as configurações utilizadas nesta etapa para a geração dos resumos candidatos nas tarefas de sumarização monodocumento e multidocumento.

Quadro 12 – Configurações adotadas para a geração dos resumos candidatos.

Tarefa	**Configurações**	**Valores**
Monodocumento	Processo de Sumarização	Abordagem baseada em conceitos usando PLI apresentada no Capítulo 4
	Representação	Sentenças, unigramas, bigramas, entidades nomeadas e dependências sintáticas rotuladas e genéricas
	Ponderação	Os treze métodos investigados no Apêndice A, o método ponderado proposto (ver 4.4) e as técnicas de pontuação de sentenças analisadas no Capítulo 3
	Outras Soluções	Inclusão da restrição apresentada na Equação 6.1 durante a execução da melhor configuração (bigramas, método ponderado, considerando o grafo de entidades) para considerar as dez melhores soluções geradas pelo modelo de PLI
	Outras	Adoção ou não do Grafo de Entidade e das restrições de coesão (ver Capítulo 4)
Multidocumento	Processo de Sumarização	Abordagem baseada em conceitos usando PLI apresentada no Capítulo 5
	Representação	Sentenças, unigramas, bigramas, entidades nomeadas e dependências sintáticas rotuladas e genéricas
	Ponderação	Os cinco métodos estatísticos investigados no Capítulo 5, o método ponderado proposto (ver 5.2) e as técnicas de pontuação de sentenças analisadas no Capítulo 3
	Outras Soluções	Adotou-se a restrição apresentada na Equação 6.1 durante a execução da melhor configuração (bigramas, método ponderado, $\lambda = 0,1$ e $\gamma = 0,6$) para considerar as quinze melhores soluções geradas pelo modelo de PLI
	Outras	Variações dos limiares de similaridade mínima entre as sentenças λ e do tamanho mínimo do grupo de sentenças γ

Ao final desta etapa, um conjunto de resumos candidatos é gerado para a coleção de documentos de entrada D. Esses candidatos refletem diferentes possíveis resumos contendo as informações mais relevantes do(s) documento(s) a serem resumido(s).

6.1.2 Seleção do resumo mais informativo

Nesta etapa, cada resumo candidato gerado anteriormente é analisado, aplicando um algoritmo de regressão para estimar a sua informatividade (cobertura de informações re-

levantes). O candidato identificado como o mais informativo é selecionado como o resumo mais adequado para representar as informações mais relevantes do(s) documento(s) de entrada. Esse processo de seleção é resumido no Algoritmo 4.

Algoritmo 4: Etapa de seleção do resumo candidato mais informativo.

Entrada: O conjunto de documentos D.
Entrada: O conjunto de resumos candidatos $resumosCandidatos$.
Entrada: O algoritmo de regressão adotado $regressor$.
Saída: O resumo candidato estimado como mais informativo $resumo_{saida}$

```
1  Início
2      maxCoberturaROUGE_1 = 0
3      para todo candidato ∈ resumosCandidatos faça
4          regressor.extrairCaracteristicas(candidato, D)
5          coberturaROUGE_1 = regressor.estimarInformatividade(candidato)
6          se coberturaROUGE_1 > maxCoberturaROUGE_1 então
7              maxCoberturaROUGE_1 = coberturaROUGE_1
8              resumo_saida = candidato
```

A criação do modelo de regressão aplicado nesta etapa requer a disponibilidade de um conjunto de treinamento, contendo um documento ou uma coleção de documentos com seus respectivos resumos. Cada resumo é representado por um conjunto de características, incluindo a medida de cobertura do ROUGE-1 (R-1) computada entre o resumo e um ou mais resumos de referência. A medida do R-1 é adotada aqui como escore de informatividade devido aos bons resultados obtidos por Hong, Marcus e Nenkova (2015) usando essa medida durante a fase de treinamento. Assim, dado um novo resumo, sem sua medida do R-1, o objetivo do algoritmo de regressão é aprender uma função que a estime.

Selecionar características adequadas para a construção do modelo de regressão é fundamental para o bom desempenho da abordagem proposta. Os trabalhos de Hong, Marcus e Nenkova (2015), Wan et al. (2015) apresentaram diversas características em nível de resumo, sentença e n-gramas, baseadas em tradicionais indicadores de relevância como posição e frequência, para a sumarização multidocumento. Alguns dos atributos mais importantes adotados em ambos os trabalhos são baseados na similaridade e divergência do conteúdo presente no resumo e na coleção de documentos de entrada. Nesses trabalhos, a semelhança foi mensurada adotando a medida de similaridade do cosseno, e a divergência foi computada usando as medidas de Kullback-Leibler (KL) (KULLBACK, 1959) e Jensen-Shannon (JS) (LIN, 2006) para comparar as distribuições de probabilidade dos unigramas e/ou bigramas presentes na coleção de documentos e no resumo.

Este trabalho propõe computar a similaridade e a divergência entre o conteúdo do resumo e do(s) documento(s) de entrada, comparando diferentes listas de conceitos, com seus respectivos pesos produzidos pelos métodos de ponderação de conceitos. Nos experimentos realizados neste trabalho, observou-se que métodos de ponderação individuais e

combinados baseados nos aspectos de posição e centralidade, demonstraram obter melhor capacidade de capturar a cobertura de informações relevantes presentes em um resumo, do que usando probabilidade, como nos trabalhos de Hong, Marcus e Nenkova (2015), Wan et al. (2015). Por isso, este trabalho de doutorado parte da hipótese de que utilizar os pesos dos conceitos em conjunto com medidas de similaridade e divergência melhora o processo de estimação da informatividade de um resumo.

As características de similaridade e divergência baseadas nos pesos dos conceitos propostas neste capítulo visam refletir a noção de que resumos informativos tendem a manter o subconjunto de conceitos mais relevantes. Para ilustrar essa suposição, imagine um documento ou conjunto de documentos $D = \{c_1 = 1.0, c_2 = 0.8, c_3 = 0.6, c_4 = 0.4, c_5 = 0.2, c_6 = 0.1\}$, representado como uma lista contendo seis conceitos, com seus respectivos valores de importância. Dado dois resumos $RC_1 = \{c_1 = 1.0, c_2 = 0.0, c_3 = 0.6, c_4 = 0.0, c_5 = 0.0, c_6 = 0.1\}$ e $RC_2 = \{c_1 = 0.0, c_2 = 0.8, c_3 = 0.0, c_4 = 0.4, c_5 = 0.2, c_6 = 0.1\}$, cada um deles também representado com a lista dos conceitos que eles mantiveram de D, e seus pesos ou zero caso contrário. Baseado na suposição acima, RC_1 deve ser mais informativo do que RC_2 porque manteve o subconjunto dos conceitos mais relevantes.

Duas formas de representação e quatro métodos de ponderação de conceitos são considerados nesta etapa para computar os atributos usados para a construção do modelo de regressão. Essas configurações foram escolhidas porque apresentaram bons resultados nos experimentos realizados nos capítulos anteriores. As duas formas de representação de conceitos consideradas nesta etapa são unigramas e bigramas, separadamente. Já os métodos de ponderação de conceitos adotados nesta etapa são: **(i)** frequência das sentenças (monodocumento) ou frequência dos documentos (multidocumento); **(ii)** posição das sentenças; **(iii)** probabilidade dos conceitos; e **(iv)** o método de ponderação combinado proposto no Capítulo 4 caso a sumarização seja monodocumento, ou no Capítulo 5 caso a tarefa em questão seja a multidocumento. Dessa forma, conforme exemplificado no parágrafo anterior, o(s) documento(s) de entrada e seus resumos candidato são representados por oito listas de conceitos com seus escores de relevância (forma de representação x método de ponderação).

6.1.3 Características usadas para a construção do modelo de regressão

No total, cem características são adotadas para construir o modelo de regressão responsável por atribuir a cada resumo candidato uma pontuação representando uma estimativa de sua medida de cobertura do ROUGE-1. As características adotadas e propostas neste trabalho são estruturadas em quatro grupos, descritos a seguir:

Importância dos Conceitos. Resumos informativos são aqueles que contém a maior quantidade de conceitos relevantes do(s) documento(s) de entrada (GILLICK et al., 2009). Baseado nessa noção, os atributos deste grupo consideram a média dos escores de re-

levância dos conceitos presentes no resumo. No total, considerando as duas formas de representação e os quatro métodos de ponderação, oito atributos são gerados.

Medidas de Similaridade. Este grupo de características computa a similaridade do resumo candidato com o(s) documento(s) de entrada, e também com outros resumos candidatos para a mesma fonte. A suposição é que um resumo informativo tende a ter um alto grau de semelhança com o(s) documento(s) originais, e é gerado por diversas configurações da abordagem de sumarização. As medidas de similaridade do cosseno e a distância euclidiana são computadas entre o resumo candidato e **(i)** o(s) documento(s) de entrada; e **(ii)** outros resumos candidatos da mesma fonte. Essas medidas foram computadas considerando somente os unigramas durante a comparação.

Além disso, computa-se a sobreposição (ver Equação 6.2), correlação de *Pearson* (ver Equação 6.3) e o coeficiente de correlação de postos de *Spearman*, entre cada uma das listas de conceitos e seus pesos do resumo e do(s) documento(s) de entrada. Essas medidas visam refletir a noção de que resumos mantendo os conceitos mais relevantes do(s) documento(s) de entrada tendem a serem mais informativos. No total, vinte e oito atributos de similaridade são gerados aqui.

$$SobreP(D, R) = \sum_{c_i \in D} Min(D_{c_i}, R_{c_i}) \tag{6.2}$$

$$Pearson(D, R) = \frac{\sum_{c_i \in D}(D_{c_i} - \mu_D)(R_{c_i} - \mu_R)}{\sqrt{\sum_{c_i \in D}(D_{c_i} - \mu_D)^2}\sqrt{\sum_{c_i \in D}(R_{c_i} - \mu_R)^2}} \tag{6.3}$$

na qual,

- D é a lista de conceitos e seus pesos (relevância) do(s) documento(s) de entrada;
- R é a lista de conceitos e seus pesos presentes no resumo;
- D_{c_i} é o peso do conceito c_i em D;
- R_{c_i} é o peso do conceito c_i em R;;
- μ_D e μ_R são as médias dos pesos dos conceitos presentes no(s) documento(s) D e no resumo candidato R, respectivamente.

O coeficiente de correlação de postos de *Spearman* é computado convertendo a lista dos pesos dos conceitos do(s) documento(s) D e do resumo candidato R em dois ranques R_d e R_R, respectivamente. Posteriormente, a medida de correlação de *Pearson* é aplicada nos ranques R_d e R_R.

Medidas de Distância e Divergência. Este grupo de atributos reflete a divergência/distância entre a lista de conceitos (com seus pesos) do resumo e do(s) documento(s) de entrada. Hong, Marcus e Nenkova (2015) observou que resumos informativos tendem a ter uma baixa divergência com a coleção de documentos de origem. Neste trabalho, são adotadas as tradicionais medidas de divergência de Kullback-Leibler (KL) (KULLBACK,

1959) (ver Equação 6.4) e Jensen-Shannon (JS) (ver Equação 6.5) investigadas em (HONG; MARCUS; NENKOVA, 2015; WAN et al., 2015). Além dessas medidas, outras medidas de divergência e distância são consideradas, buscando melhor descrever o problema, sendo elas: *Skew divergence* (LEE, 1999) (ver Equação 6.6), Entropia Cruzada (ver Equação 6.7), Distância Média (DM) (ver Equação 6.8) e Coeficiente de Divergência (CD) (ver Equação 6.9) (SHIRKHORSHIDI; AGHABOZORGI; WAH, 2015). As medidas de divergência/distância buscam refletir a noção de que resumos informativos tendem a ter uma baixa divergência/distância com o(s) documento(s) de entrada, pois mantêm os conceitos mais relevantes. No total, quarenta e oito atributos de divergência são produzidos neste grupo.

As medidas de divergência Kullback-Leibler (KL), Jensen-Shannon (JS), *Skew divergence* e Entropia Cruzada são originalmente aplicadas para medir a distância entre uma distribuição de probabilidade inicial P e uma segunda distribuição esperada Q. Analisando-as, observou-se que a divergência entre P e Q aumenta a medida que elementos com maior probabilidade em P estão ausentes ou com uma probabilidade menor em Q. Esse comportamento é similar ao que acreditamos que acontece com resumos informativos, ou seja, eles tendem a manter os conceitos mais relevantes do(s) documento(s) de entrada. Dessa forma, supõe-se que essas medidas podem ser aplicadas para mensurar a divergência entre a lista de conceitos do(s) documento(s) de entrada e do resumo candidato.

Dado a lista de conceitos, com seus respectivos pesos, do(s) documento(s) de entrada D e do resumo candidato R, sendo ambos normalizados de forma a garantir que $\sum_{c_i \in D} D_{c_i} = 1$ e $\sum_{c_i \in R} R_{c_i} = 1$, as medidas de divergência/distância são computadas conforme apresentado a seguir:

- **Kullback-Leibler** (KULLBACK, 1959): essa medida clássica de distância é usada aqui para computar o quão divergente a lista de pesos dos conceitos do(s) documento(s) de entrada é, em relação aos pesos dos conceitos remanescente no resumo candidato. Essa medida é indefinida quando $D_{c_i} > 0$ e $R_{c_i} = 0$, ou seja, quando o conceito c_i não está presente no resumo candidato. Como esse cenário é frequente, utilizou-se um fator de suavização $R_{c_i} = 0.0001$ conforme sugerido por (LOUIS; NENKOVA, 2009).

$$KL(D||R) = \sum_{c_i \in D} D_{c_i} \times log\left(\frac{D_{c_i}}{R_{c_i}}\right) \tag{6.4}$$

- **Jensen-Shannon** (LIN, 2006): essa medida é baseada na Kullback-Leibler e incorpora a ideia de que se as duas listas de pesos dos conceitos D e R são similares, elas também devem ser semelhantes a uma terceira lista M representando a média das

distâncias entre D e R.

$$JS(D||R) = \frac{KL(D||M) + KL(R||M)}{2} \tag{6.5}$$

- **Skew divergence** (LEE, 1999): essa medida também é baseada na Kullback-Leibler, e utiliza uma estratégia de suavização que combina as duas listas de pesos D e R, em um grau determinado pelo parâmetro $\alpha \in [0, 1]$. Observou-se que definido um valor mais próximo de zero $\alpha = 0.1$, um maior peso é dado para os casos em que o resumo R manteve uma ocorrência de um conceito de D.

$$Skew(D||R) = KL(D||(D \times \alpha + R \times (1 - \alpha))) \tag{6.6}$$

- **Entropia Cruzada (EC):** é similar a medida Kullback-Leibler, mas sem a necessidade de um fator de suavização.

$$EC(D, R) = \sum_{c_i \in D} D_{c_i} \times logR_{c_i} \tag{6.7}$$

- **Distância Média (DM)** e **Coeficiente de Divergência (CoefDiv)** (SHIRKHORSHIDI; AGHABOZORGI; WAH, 2015): são duas medidas de distância comumente usadas em conjunto com algoritmos de agrupamento para medir a dissimilaridade entre dois vetores de atributos contínuos.

$$DM(D, R) = \frac{1}{|D_{c_i}|} \times \left(\sum_{c_i \in D} (D_{c_i} - R_{c_i})^2 \right)^{\frac{1}{2}} \tag{6.8}$$

$$CoefDiv(D, R) = \left(\frac{1}{|D_{c_i}|} \times \sum_{c_i \in D} \left(\frac{D_{c_i} - R_{c_i}}{D_{c_i} + R_{c_i}} \right)^2 \right)^{\frac{1}{2}} \tag{6.9}$$

na qual,

- D é a lista de pesos (relevância) dos conceitos do(s) documento(s) de entrada;
- R é a lista de pesos dos conceitos presentes no resumo candidato;
- $|D_{c_i}|$ é o total de conceitos presentes em D;
- D_{c_i} e R_{c_i} são os escores de importância do conceito c_i em D e R, respectivamente. Vale ressaltar que D_{c_i} e R_{c_i} podem ser diferentes por conta do processo de normalização;
- M é o vetor médio entre D e R, no qual $M_{c_i} = \frac{D_{c_i} + R_{c_i}}{2}$.

Outras Características. As outras dezesseis características adotadas são:

- **Pontuação do grafo de entidades**: A média dos escores atribuído pelo modelo de grafo de entidades (ver Equação 4.5) a cada sentença presente no resumo.
- **Probabilidades Multinomiais dos n-gramas** (HONG; MARCUS; NENKOVA, 2015): A média das probabilidades dos n-gramas (unigramas e bigramas, separadamente) presentes no resumo normalizado pelo número total de n-gramas no resumo.
- **Proporção de palavras na primeira sentença** (HONG; MARCUS; NENKOVA, 2015): A proporção de palavras presente no resumo que estão na primeira sentença de cada documento em $d_i \in D$.
- **Proporção de substantivos, adjetivos, advérbios e verbos** (WAN et al., 2015): A proporção de substantivos, adjetivos, advérbios e verbos, em relação ao total de palavras do resumo.
- **Proporção de palavras únicas** (WAN et al., 2015): A proporção de palavras distintas no resumo, em relação ao total de palavras do resumo.
- **Proporção de n-gramas**: A proporção de n-gramas (unigramas e bigramas, separadamente) distintas no resumo, em relação ao total de n-gramas da coleção de documentos D.
- **Proporção de Palavras no(s) Título(s)**: A proporção de palavras presente no resumo que estão no título de cada documento em $d_i \in D$.
- **Posição das sentenças** (WAN et al., 2015): A média dos escores do método de posição das sentenças (ver Equação 4.3) presentes no resumo.
- **Tamanho das sentenças** (WAN et al., 2015): O total de palavras da maior e menor sentença, além do tamanho médio das sentenças presentes no resumo.
- **Confiança do método de sumarização**: Estes atributos usam como escore de confiança, as médias das medidas de cobertura do ROUGE-1 e ROUGE-2 obtidos por resumos gerados anteriormente pela configuração usada para produzir o resumo em análise. Os escores de confiança são derivados do conjunto de treinamento. O pressuposto é que resumos produzidos por configurações com um histórico de gerar resumos informativos, são mais prováveis de também serem informativos.
- **Total de sentenças** (WAN et al., 2015): O total de sentenças contidas no resumo.

6.2 Experimentos

Esta seção apresenta e discute os experimentos realizados para avaliar diferentes aspectos da abordagem proposta, nas tarefas de sumarização monodocumento e multidocumento. Diversos experimentos foram conduzidos buscando avaliar os seguintes aspectos:

(i) uma análise individual das características adotadas para criar o modelo de regressão proposto (ver Seção 6.2.2); **(ii)** uma avaliação comparativa do desempenho de cinco algoritmos de regressão, aplicados para estimar a medida de cobertura do ROUGE-1 de cada resumo candidato (ver Seção 6.2.3); e **(iii)** uma comparação entre a abordagem proposta e diversos sistemas do estado da arte (ver Seção 6.2.4).

Antes de discutir os resultados obtidos, uma breve descrição do ambiente experimental adotado é apresentada na próxima seção.

6.2.1 Configurações dos Experimentos

Corpora. Diversos experimentos foram realizados neste capítulo, visando avaliar a abordagem proposta nas tarefas de sumarização genérica monodocumento e multidocumento, no domínio de artigos de notícias escritos em Inglês. Na sumarização monodocumento foram adotados os corpora do DUC 2001-2002 e CNN. Já na sumarização multidocumento os corpora do DUC 2001-2004 foram usados. Na Tabela 27 são apresentadas a seguintes estatísticas de cada corpus, geradas usando a ferramenta *Stanford CoreNLP* (MANNING et al., 2014): **(i)** Total de grupos de documentos; **(ii)** Total de sentenças do corpus (com repetição); **(iii)** Total de palavras do corpus (com repetição); **(iv)** As tarefas de sumarização nas quais o corpus são adotados; e **(v)** Total de resumos candidatos gerados para cada documento (monodocumento) ou grupo de documentos (multidocumento).

Tabela 27 – Estatísticas dos corpora do CNN e do DUC adotados nos experimentos.

Corpus	#Grupos	#Documentos	#Sentenças	#Palavras	Tarefas	#Candidatos
CNN	0	3.000	115.649	2.628.336	Mono	60
DUC 2001	30	308	11.026	269.990	Mono e Multi	100
DUC 2002	59	533	14.370	348.012	Mono e Multi	(60) 100
DUC 2003	30	298	7,691	197.483	Multi	100
DUC 2004	50	500	13.135	336.073	Multi	100

Medidas de avaliação. As medidas de cobertura do ROUGE-1 (R-1) e ROUGE-2 (R-2) foram adotadas neste capítulo para avaliar os resumos gerados automaticamente. Os escores produzidos pela ferramenta ROUGE (LIN, 2004) foram as medidas oficiais de avaliação durante as competições do DUC de 2001 até 2004. As medidas de cobertura do R-1 e R-2 foram adotadas porque elas apresentaram uma forte correlação com avaliações conduzidas por humanos (LIN, 2004; OWCZARZAK et al., 2012). A versão 1.5.5 da ferramenta ROUGE foi usada, realizando *stemming* e sem a remoção de *stopwords*. Para isso, a seguinte linha de comando foi adotada: *-m -f A*.

Tamanho dos resumos. Nos corpora do DUC, em ambas as tarefas (monodocumento e multidocumento), o parâmetro "-l 100" foi usado para definir que a ferramenta do ROUGE só deve considerar as cem primeiras palavras do resumo durante a avaliação. O

parâmetro do tamanho máximo do resumo a ser gerado L no modelo de PLI (ver Equação 5.3) foi definido com o valor máximo de 105 palavras. Esse limiar foi adotado dessa forma para permitir a geração de resumos com tamanhos compatíveis com os produzidos pelos trabalhos relacionados, ou seja, em torno de 100 palavras. Em experimentos anteriores observou-se que ao definir o parâmetro L do modelo de PLI com 100 palavras, os resumos gerados eram muito menores do que os gerados pelos trabalhos relacionados. No corpus CNN, a taxa de compressão usada foi de 10% do total de sentenças do documento de entrada.

Quantidade de resumos candidatos gerados. No total, cem resumos candidatos são gerados para cada documento ou coleção de documentos de entrada para a sumarização monodocumento e multidocumento, respectivamente. Esses resumos candidatos são gerados usando as configurações apresentadas no Quadro 12. Os corpora CNN e DUC 2002, para a sumarização monodocumento, possuem uma grande quantidade de documentos. Para evitar um alto custo computacional durante a etapa de treinamento e teste, nesses dois corpora são gerados somente sessenta resumos candidatos para cada documento. Para isso, as representações de conceitos usando entidades nomeadas e dependências sintáticas rotuladas e genéricas, não foram consideradas durante a etapa de geração dos resumos candidatos.

6.2.2 Avaliação individual das características

Este primeiro experimento analisa, individualmente, as características adotadas para a construção do modelo de regressão usado para estimar a medida de cobertura do ROUGE-1 dos resumos candidatos. Para isso, a correlação de *Pearson* e *Spearman* foram computadas entre cada uma das cem características adotadas e os escores do R-1 dos resumos candidatos gerados. Por questões de espaço, somente as vinte características com maior correlação de *Pearson* são apresentadas.

As seções a seguir apresentam os resultados obtidos neste experimento, nas tarefas de sumarização monodocumento e multidocumento, respectivamente.

Monodocumento

A Tabela 28 apresenta os resultados da correlação de *Pearson* (P) e *Spearman* (S) entre a medida de cobertura do ROUGE-1 e as características usadas para a construção do modelo de regressão na tarefa de sumarização monodocumento. Para facilitar o entendimento dos atributos de similaridade e divergência apresentados, adotamos um padrão de nomenclatura no seguinte formato: Nome da medida de similaridade/divergência adotada, mais o nome da lista de pesos dos conceitos no qual a medida foi aplicada (em Itálico). Por exemplo, o nome Dist. Média *Bigrama Pos. Sent.* indica o atributo gerado aplicando a

Tabela 28 – Os vinte atributos com maior correlação de *Pearson* (P) e *Spearman* (S) com a medida de cobertura do R-1 na tarefa de sumarização monodocumento. Os atributos com maior correlação em cada corpus são destacados em negrito.

Característica	**CNN**		**DUC 2001**		**DUC 2002**	
	P	S	P	S	P	S
Dist. Média *Bigrama Pos. Sent.*	-0,5727	-0,4880	-0,4060	-0,3647	-0,4998	-0,4166
Dist. Média *Bigrama Método Comb.*	-0,5793	-0,4985	-0,4205	-0,3839	-0,5168	-0,4302
CoefDiv *Bigrama Pos. Sent.*	-0,5757	-0,5013	-0,4474	-0,3997	-0,5228	-0,4199
CoefDiv *Bigrama Método Comb.*	-0,5822	-0,5057	-0,4628	-0,4118	-0,5304	-0,4259
Entropia Cruzada *Bigrama Freq. Sent.*	-0,5592	-0,4594	-0,3735	-0,3224	-0,4318	-0,3240
Entropia Cruzada *Bigrama Pos. Sent.*	-0,5687	-0,4784	-0,4470	-0,4066	-0,5166	-0,4390
Entropia Cruzada *Unigrama Método Comb.*	-0,5677	-0,4729	-0,4161	-0,3626	-0,4972	-0,4064
JS *Bigrama Pos. Sent.*	-0,5820	-0,4851	-0,4453	-0,4058	-0,5246	-0,4406
JS *Bigrama Método Comb.*	-0,5915	-0,4958	-0,4638	-0,4285	-0,5373	-0,4494
JS *Unigrama Método Comb.*	-0,5791	-0,4786	-0,4156	-0,3584	-0,5014	-0,4072
KL *Bigrama Pos. Sent.*	-0,5687	-0,4784	-0,4470	-0,4066	-0,5166	-0,4390
KL *Bigrama Método Comb.*	-0,5794	-0,4898	-0,4659	**-0,4305**	-0,5299	-0,4450
KL *Unigrama Método Comb.*	-0,5677	-0,4729	-0,4161	-0,3626	-0,4972	-0,4064
SobreP, *Bigrama Freq. Sent.*	0,5466	0,4561	0,3651	0,3181	0,4179	0,3220
SobreP, *Bigrama Pos. Sent.*	0,5683	0,4884	0,4444	0,4071	0,5107	0,4413
Skew *Bigrama Pos. Sent.*	-0,5820	-0,4882	-0,4463	-0,4074	-0,5239	-0,4412
Skew *Bigrama Método Comb.*	**-0,5917**	-0,4986	-0,4651	-0,4278	-0,5368	-0,4503
Skew *Unigrama Método Comb.*	-0,5789	-0,4800	-0,4151	-0,3581	-0,4995	-0,4075
Sim. Cosseno Outros Candidatos	0,5668	0,4929	**0,4707**	0,4259	0,5374	0,4668
Dist. Euclideana Outros Candidatos	0,5778	**0,5073**	0,4581	0,4111	**0,5383**	**0,4690**

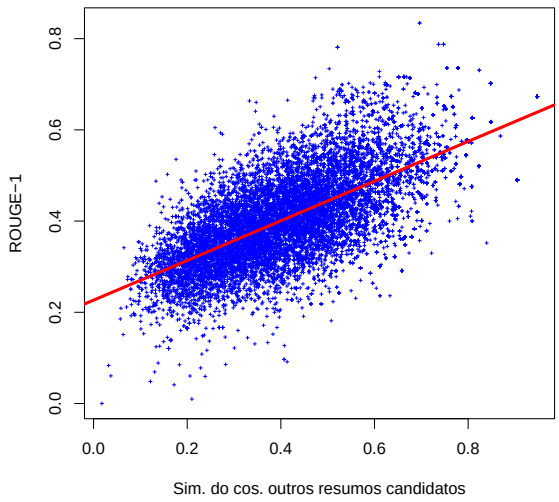
medida de distância média (ver Equação 6.8) para comparar as listas de pesos produzidas usando bigramas como conceitos e posição de sentenças como método de ponderação.

Os grupos de características que capturam a similaridade e a divergência entre o(s) documento(s) e o resumo candidato apresentam as maiores correlações com a medida de cobertura do R-1 nos três corpora. As correlações variam de fraca ($0,3 \leq |correlacao| < 0,5$) a moderada ($0,5 \leq |correlacao| < 0,7$), dependendo do atributo e do corpus adotado. Nenhum dos atributos investigados apresentou uma correlação forte ($0,7 \leq |correlacao|$) com a medida do R-1 neste experimento. Tal comportamento é esperado dada a complexidade e a subjetividade envolvida no processo de sumarização. Contudo, podemos observar que alguns dos atributos adotados demonstram ser bons indicadores para identificar se um resumo é informativo. Os oitenta atributos omitidos apresentam correlações variando de insignificante ($0 \leq |correlacao| < 0,3$) a moderada.

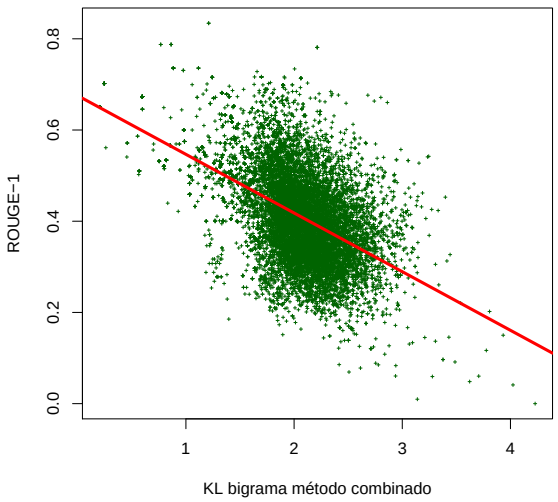
As características de similaridade e divergência propostas neste trabalho, considerando os métodos de ponderação de conceitos da frequência e posição das sentenças, além do método combinado, apresentaram correlações mais fortes com o R-1 do que os atributos usando probabilidade como adotado em (HONG; MARCUS; NENKOVA, 2015). Esses resultados demonstram que esses métodos de ponderação de conceitos apresentam melhor desempenho na identificação de informações relevantes que devem estar presentes em um resumo informativo.

A Figura 11 apresenta graficamente a correlação entre a medida de cobertura do R-1 e as quatro características com maior correlação no DUC 2001, sendo elas: Similaridade do cosseno com outros resumos candidatos (SimCos), e as medidas de divergência KL, JS e Skew usando a lista de pesos adotando bigramas e o método de ponderação combinado. A Figura 11a demonstra que o atributo SimCos possui uma correlação moderada positiva com a medida do R-1, ou seja, resumos contendo maior valor nesse atributo, em geral, também possuem maior escore na medida do R-1. A relação inversa (correlação negativa) acontece com as medidas de divergência (Figura 11b, Figura 11c e Figura 11d), ou seja, resumos com baixo valor de divergência, em geral, possuem maior valor de cobertura no R-1.

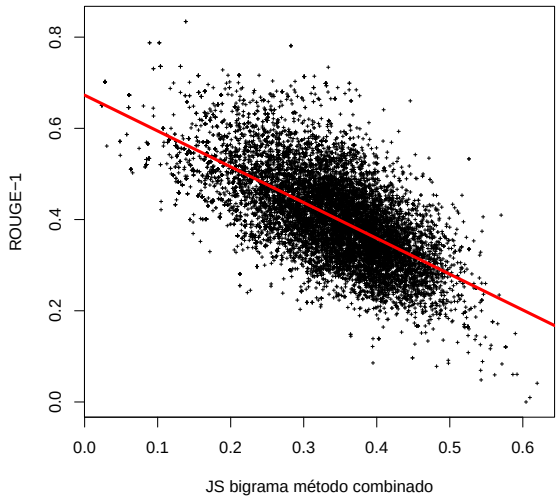
Em resumo, os resultados obtidos demonstram que resumos informativos tendem a: **(i)** manter a maior quantidade de conceitos relevantes, ou seja, possuem uma alta similaridade e uma baixa divergência com o documento de entrada, principalmente considerando os atributos gerados aplicando os métodos de posição das sentenças e o método de ponderação combinado proposto; e **(ii)** possuem uma alta similaridade com os outros resumos candidatos gerados para o documento de entrada aplicando diferentes configurações da abordagem de sumarização monodocumento.



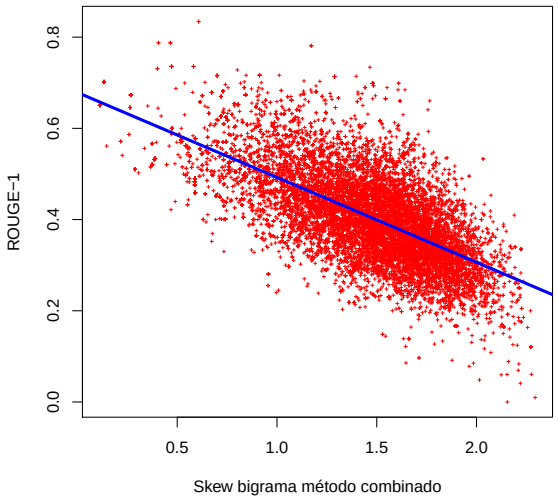
(a) R-1 vs Sim. do cosseno com outros resumos candidatos.



(b) R-1 vs KL bigrama método combinado.



(c) R-1 vs JS bigrama método combinado.



(d) R-1 vs Skew bigrama método combinado.

Figura 11 – Quatro características com maior correlação de *Pearson* com a medida de cobertura do R-1 no corpus do DUC 2001, considerando somente os resumos candidatos gerados adotando bigramas como conceitos.

Multidocumento

Na Tabela 29 são apresentas as correlações de *Pearson* (P) e *Spearman* (S) das vinte características com maior média de correlação de *Pearson* com a medida de cobertura do R-1, na tarefa de sumarização multidocumento adotando os quatro corpora do DUC 2001-2004. O mesmo padrão de nomenclatura das características avaliadas adotado no experimento monodocumento é mantido aqui.

Os grupos de características que capturam a relevância, a confiança da configuração

Tabela 29 – Os vinte atributos com maior correlação de *Pearson* (P) e *Spearman* (S) com a medida de cobertura do R-1 na tarefa de sumarização multidocumento. Os atributos com maior correlação em cada corpus são destacados em negrito.

Característica	**DUC 2001**		**DUC 2002**		**DUC 2003**		**DUC 2004**	
	P	S	P	S	P	S	P	S
EC Bigrama *Pos. Sent.*	-0,5742	-0,4897	-0,5934	0,2336	-0,5249	-0,3970	-0,6457	-0,5317
EC Bigrama *Método Comb.*	-0,5565	-0,4857	-0,5613	0,2421	-0,5620	-0,4406	-0,6652	**-0,5547**
JS Bigrama *Pos. Sent.*	-0,5703	-0,4774	-0,5836	0,2364	-0,5254	-0,3949	-0,6396	-0,5234
JS Bigrama *Método Comb.*	-0,5716	-0,4960	-0,5744	0,2435	-0,5660	-0,4368	-0,672	-0,5539
KL Bigrama *Pos. Sent.*	-0,5742	-0,4897	-0,5934	0,2336	-0,5249	-0,3970	-0,6457	-0,5317
KL Bigrama *Método Comb.*	-0,5565	-0,4857	-0,5613	0,2421	-0,5620	-0,4406	-0,6652	**-0,5547**
SobreP. Bigrama *Pos. Sent.*	0,5600	0,4669	0,5664	0,2422	0,5107	0,3928	0,6249	0,5227
SobreP. Bigrama *Método Comb.*	0,5615	0,4968	0,5574	0,2463	0,5475	0,4342	0,6533	0,5541
Pearson Bigrama *Freq. Doc.*	0,5629	0,4845	0,5200	**0,2663**	0,5530	0,4520	0,6541	0,52
Pearson Bigrama *Pos. Sent.*	0,5601	0,4824	0,5909	0,2237	0,5472	0,4130	0,6735	0,5503
Pearson Bigrama *Método Comb.*	0,5127	0,4412	0,5277	0,2437	0,5455	0,4354	0,6587	0,5327
Pearson Unigrama *Pos. Sent.*	0,5154	0,4029	0,5584	0,2346	0,5653	0,4382	0,6317	0,4561
Skew Bigrama *Pos. Sent.*	-0,5656	-0,4670	-0,5747	0,2396	-0,5210	-0,3929	-0,633	-0,5227
Skew Bigrama *Método Comb.*	-0,5718	-0,4969	-0,5720	0,2450	-0,5620	-0,4340	-0,6686	-0,5541
Sim. Cosseno Outros Candidatos	0,5661	**0,5033**	0,5783	0,2357	0,5594	0,4552	0,65	0,5436
Confiança ROUGE-1	0,5721	0,4554	0,6062	0,1725	0,5265	0,4072	0,664	0,5084
Confiança ROUGE-2	0,5350	0,4410	0,5813	0,1952	0,5134	0,4209	0,6525	0,5139
Unigrama *Freq. Doc.*	0,5767	0,4379	0,5838	0,2296	**0,5774**	**0,4796**	0,6689	0,5039
Unigrama *Pos. Sent.*	**0,5924**	0,4510	**0,6230**	0,2126	0,5563	0,4280	**0,6859**	0,541
Unigrama *Método Comb.*	0,5580	0,4380	0,5809	0,2283	0,5602	0,4324	0,6762	0,5326

usada para gerar o resumo, e a similaridade/divergência entre a coleção de documentos e o resumo candidato apresentaram as maiores correlações com a medida de cobertura do R-1. Assim como na sumarização monodocumento, as correlações dos vinte atributos variam de fraca a moderada, dependendo do corpus considerado.

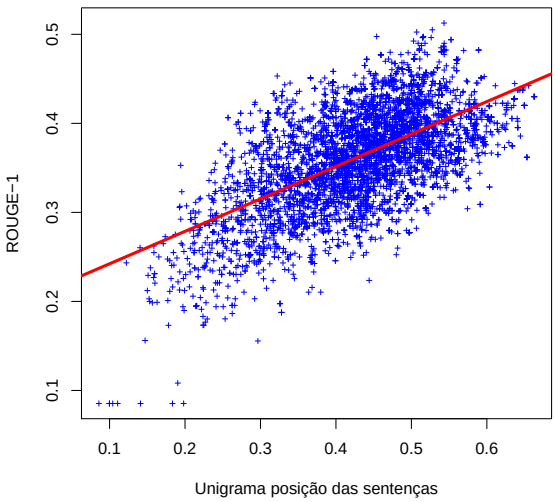
Uma maior diversidade de atributos pode ser observada na Tabela 29 em comparação com os resultados da sumarização monodocumento. Os atributos de divergência baseados nas medidas de Entropia Cruzada (EC), Kullback-Leibler (KL), Jensen-Shannon (JS) e *Skew*, considerando bigramas ponderados pelos métodos de posição de sentenças e o método combinado, também obtiveram boa correlação para a sumarização multidocumento. Além dessas características, outras que também apresentaram boas correlações em ambos os experimentos são: Similaridade do cosseno com outros candidatos, Sobreposição considerando bigrama e o método de posição das sentenças e o método de ponderação combinado.

Diferentemente da sumarização monodocumento, os seguintes atributos apresentaram bom desempenho para sumarização multidocumento: **(i)** Os atributos de relevância considerando unigramas e os métodos posição das sentenças, frequência dos documento, e o método ponderado combinado; **(ii)** As características geradas usando a medida de correlação de *Pearson* adotando os métodos posição das sentenças (unigrama e bigrama), frequência dos documento (bigrama), e o método ponderado combinado (bigrama); **iii** Os atributos de confiança que derivam do conjunto de treinamento uma média do R-1 e R-2 dos resumos gerados anteriormente adotando a configuração usada para produzir os resumos candidatos.

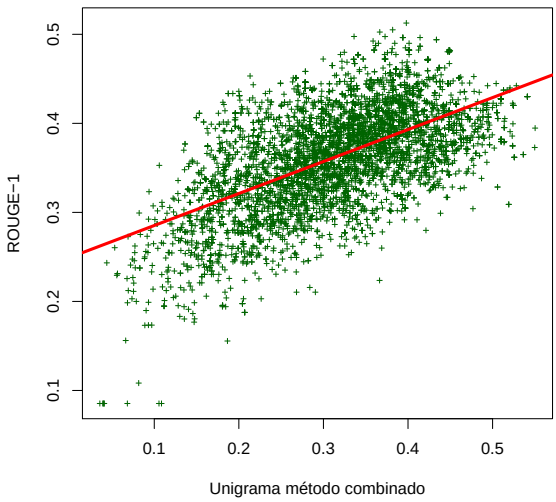
Na Figura 12 são apresentadas graficamente a correlação entre a medida de cobertura do R-1 e os quatro atributos com maior correlação de *Pearson* no corpus do DUC 2004, sendo elas: Unigrama considerando os métodos de posição das sentenças e o método de ponderação combinado, correlação de *Pearson* e a medida divergência *Jensen-Shannon* usando bigramas e o método combinado.

A Figura 12a e a Figura 12b demonstram que os atributos de relevância usando unigramas e os métodos de posição das sentenças e o método combinado, possuem uma correlação positiva com a medida do R-1. Tal comportamento indica que resumos contendo uma maior quantidade de unigramas ponderados por esses dois métodos, em geral, possuem maior escore na medida do R-1. A Figura 12c também ilustra uma correlação positiva moderada entre a medida do R-1 e o atributo gerado computando a correlação de *Pearson* entre a lista de pesos do método ponderado combinado usando bigramas, da coleção de documentos e do resumo candidato. Uma correlação moderada negativa é apresentada Figura 12d usando a medida de JS considerando bigramas e os pesos gerados pelo método ponderado combinado.

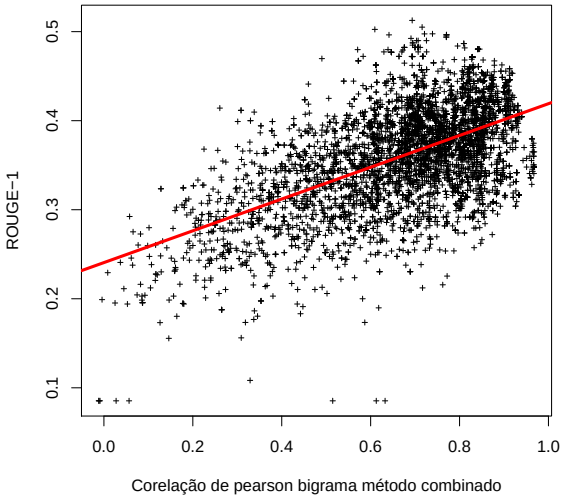
Os resultados obtidos demonstram que resumos informativos de múltiplos documentos tendem a: **(i)** possuir uma alta similaridade e uma baixa divergência com a coleção de



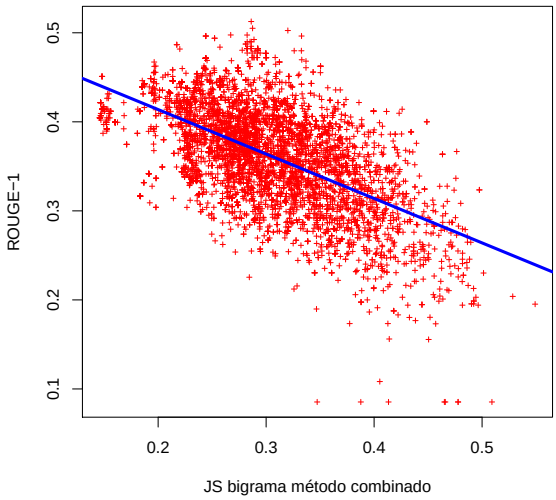
(a) R-1 vs Unigrama posição das sentenças.



(b) R-1 vs Unigrama método combinado.



(c) R-1 vs Correlação de *Pearson* bigrama método combinado.



(d) R-1 vs JS bigrama método combinado.

Figura 12 – Correlação de *Pearson* entre a medida de cobertura do R-1 e as quatro características com maior correlação no corpus do DUC 2004.

documentos de entrada, principalmente adotando os atributos gerados aplicando os métodos de posição das sentenças e o método de ponderação combinado; **(ii)** apresentar uma alta similaridade com outros resumos candidatos gerados para a coleção de documentos de entrada aplicando diferentes configurações da abordagem de sumarização multidocumento; **(iii)** conservar a maior quantidade de unigramas com alto escore de relevância ponderados pelos métodos de frequência dos documentos, posição das sentenças e o método combinado; e **(iv)** serem gerados por configurações com um histórico de produzir resumos com alto escores nas medidas de cobertura do R-1 e R-2.

6.2.3 Avaliação comparativa entre diferentes algoritmos de regressão

O objetivo deste segundo experimento é avaliar o desempenho de diferentes algoritmos de regressão para estimar a medida de cobertura do ROUGE-1 dos resumos candidatos gerados, combinando as características investigadas neste trabalho. Os seguintes algoritmos de regressão disponíveis na plataforma Weka (MANNING et al., 2014) são avaliados utilizando suas configurações padrões: Regressão linear com o método de seleção de atributos M5, Regressão linear com mínimo erro quadrático (LeastMedSq) (ROUSSEEUW, 1984), Perceptrons de múltiplas camadas com uma camada escondida (MLP), Máquina de vetores de suporte para regressão usando como núcleo uma função de base radial (SMOreg) (SHEVADE et al., 1999), e Redes de funções de base radial para regressão (Regressão RBF) (FRANK, 2014).

Monodocumento

A metodologia de validação cruzada com dez subconjuntos (*10-fold cross validation*) é adotada neste experimento para avaliar o desempenho dos cinco algoritmos de regressão para a tarefa de sumarização monodocumento. Assim, cada corpus é dividido em dez subconjuntos de documentos, e a validação cruzada é executada internamente em cada corpus. Para cada subconjunto, o processo de avaliação é executado de acordo com as duas seguintes etapas:

- **Treinamento:** Nesta etapa, os $k-1$ subconjuntos são usados para treinar o modelo de regressão.
- **Teste:** O subconjunto não selecionado na etapa de treinamento é usado como conjunto de teste. O modelo de regressão criado na etapa anterior é aplicado para estimar a medida de cobertura do R-1 dos resumos candidatos de todos os documentos pertencentes ao conjunto de teste. Para cada documento, o resumo candidato com maior valor estimado do R-1 é selecionado como o mais informativo.

O processo acima é executado dez vezes e a execução com desempenho mais próximo da média global de todos os algoritmos, com base nas medidas de cobertura do R-1 e R-2 foi selecionada. Na Tabela 30 são apresentados os resultados deste experimento considerando as medidas de: **(i)** correlação de *Pearson* (P) e *Spearman* (S); **(ii)** a média da soma dos quadrados dos residuais, do inglês, *Residual Sum of Squares* (RSS) (ver Equação 6.10); e **(iii)** a média das medidas de cobertura do R-1 e R-2 (desvio padrão entre parênteses) dos resumos selecionados como os mais informativos para cada documento de entrada.

$$\frac{\sum_{r \in R}(x - \widehat{x})^2}{|R|} \tag{6.10}$$

No qual,

- x_i é o valor da medida de cobertura do R-1 do resumo r;
- $\widehat{x_i}$ é o valor estimado pelo algoritmo de regressão da medida de cobertura do R-1 do resumo r;
- $|R|$ é o total de resumos candidatos analisados.

Tabela 30 – Resultados da análise comparativa entre os cinco algoritmos de regressão para a tarefa de sumarização monodocumento. Em negrito são destacados os algoritmos com melhor desempenho em cada medida de avaliação e corpus adotado. O símbolo † nas medidas do R-1 e R-2 indicam cenários de equivalência estatística com o algoritmo de melhor desempenho (em negrito).

Algoritmos	**CNN**				
	P	S	RSS (%)	R-1 (%)	R-2 (%)
LeastMedSq	0,635	0,535	2,16	58,42† (20,04)	41,09 (25,52)
MLP	0,632	0,530	2,18	58,46† (19,91)	41,29† (25,56)
Regressão Linear	**0,641**	**0,539**	**2,14**	**58,58** (19,87)	**41,36** (25,43)
Regressão RBF	0,635	0,536	2,15	58,36† (20,23)	41,14† (25,78)
SMOreg	0,636	0,536	2,19	58,42† (19,86)	41,17† (25,46)
Algoritmos	**DUC 2001**				
	P	S	RSS (%)	R-1 (%)	R-2 (%)
LeastMedSq	0,556	0,491	0,52	46,20† (9,59)	20,79† (11,61)
MLP	0,551	0,447	0,57	45,30 (9,84)	20,03 (11,78)
Regressão Linear	0,549	0,485	0,52	**46,37** (9,76)	**21,10** (11,71)
Regressão RBF	0,555	0,492	0,52	46,04† (9,96)	20,91† (11,91)
SMOreg	**0,558**	**0,496**	**0,51**	45,98 (9,68)	20,54 (11,64)
Algoritmos	**DUC 2002**				
	P	S	RSS (%)	R-1 (%)	R-2 (%)
LeastMedSq	0,637	0,519	0,432	**49,82** (8,37)	**23,96** (9,75)
MLP	0,629	0,506	0,450	49,50† (8,27)	23,63† (9,55)
Regressão Linear	0,637	0,519	0,432	49,78† (8,40)	23,92† (9,74)
Regressão RBF	0,638	0,522	0,432	49,58† (8,27)	23,72† (9,68)
SMOreg	**0,643**	**0,526**	**0,427**	49,67† (8,35)	23,76† (9,53)

No corpus CNN, o algoritmo de regressão linear obteve o melhor desempenho em todas as medidas de avaliação adotadas. Os algoritmos apresentaram resultados similares em todas as medidas de avaliação neste corpus. A regressão linear obteve melhor desempenho nas duas medidas do ROUGE, contudo uma diferença estatística foi observada somente comparando a medida do R-2 com o algoritmo LeastMedSq, adotando um nível de 95% de confiança (*Wilcoxon signed rank*).

No DUC 2001, o algoritmo SMOreg apresentou melhores resultados na medida de RSS e nas correlações de *Pearson* e *Spearman*. Em relação às medidas do R-1 e R-2, o algoritmo de regressão linear superou todos os outros algoritmos, mas essas melhorias só foram estatisticamente significativas quando comparadas com os resultados dos algoritmos SMOreg e MLP.

No DUC 2002, o algoritmo SMOreg mais uma vez apresentou a melhor performance nas medidas de RSS e nas correlações de *Pearson* e *Spearman*. O algoritmo LeastMedSq obteve o melhor resultado com base nas medidas do R-1 e R-2. Contudo, em ambas as medidas, nenhuma diferença estatística foi observada.

De maneira geral, os cinco algoritmos de regressão avaliados obtiveram resultados similares nas medidas de avaliação consideradas. Somente no DUC 2001, pode-se observar uma discrepância maior no desempenho dos algoritmos MLP e SMOreg, com base nas medidas do R-1 e R-2 em relação ao algoritmo com melhor resultado. Correlações mais fortes baseados nos métodos de *Pearson* e *Spearman* foram obtidas pelos algoritmos de regressão em comparação com as características individualmente, mas ainda assim, elas representam uma correlação moderada com a medida do R-1 em todos os corpora. A medida de RSS demonstra que os algoritmos, em geral, conseguem estimar valores próximos aos reais escores da medida do R-1. Somente no corpus CNN que um percentual maior de erro, em torno de 2%, foi obtido.

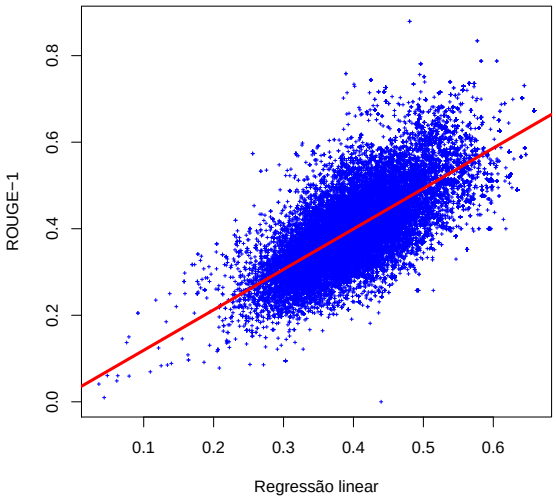
Com base nos resultados das medidas do R-1 e R-2, nos três corpora testados, é possível concluir que os algoritmos de regressão linear e LeastMedSq obtiveram as duas melhores performances globais. O algoritmo de regressão linear é considerado o mais adequado neste experimento, porque apresentou bons resultados nas medidas do ROUGE, e também requer um menor esforço computacional durante as fases de treinamento e teste, em comparação com os outros algoritmos avaliados.

Um ponto interessante observado é que nem sempre o algoritmo com maior correlação e menor erro na medida de RSS, também apresenta a melhor performance nas medidas do R-1 e R-2. Acredita-se que tal comportamento acontece porque os algoritmos podem cometer erros em situações específicas. Por exemplo, um algoritmo pode aproximar bem os valores estimados para a grande maioria dos resumos candidatos, mas nos resumos com maior valor do R-1, ele pode estimar valores um pouco distante dos reais. Esse erro impactará diretamente na escolha do resumo candidato mais informativo, fazendo com que o algoritmo selecione um resumo candidato com um valor do R-1 menor do que outros resumos candidatos existentes.

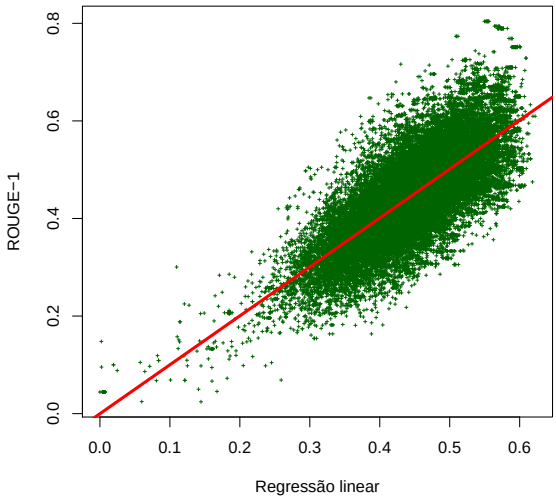
Um outro ponto observado durante este experimento é a existência de diversos resumos candidatos com baixos escores na medida do R-1, mas com características muito similares com resumos com altos valores (falso positivos). Isso acontece devido à complexidade e a subjetividade da tarefa de sumarização. Esse problema diminuiu o desempenho dos algoritmos de regressão nas medidas de avaliação adotadas.

Na Figura 13 é apresentada graficamente a correlação de *Pearson* obtida entre os valores reais e os estimados da medida do R-1, usando o algoritmo de regressão linear nos três corpora avaliados. É possível observar que a regressão linear obteve uma correlação moderada positiva em todos os corpora. Contudo, observa-se também uma dispersão nos dados, principalmente no corpus CNN. Isso indica, que existem resumos com uma

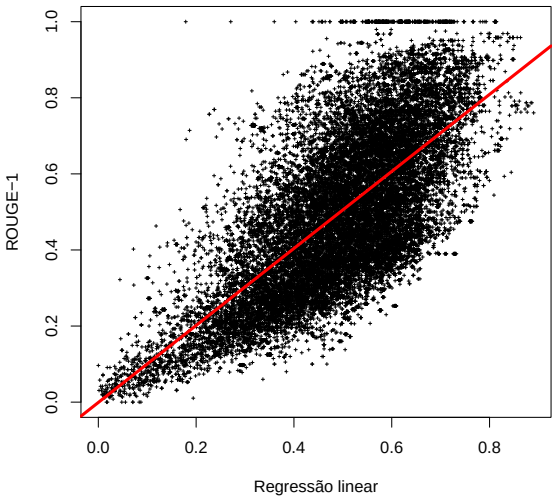
discrepância entre o valor real do R-1 e o valor estimado pela regressão linear.



(a) DUC 2001.



(b) DUC 2002.



(c) CNN.

Figura 13 – Correlação de *Pearson* entre a medida de cobertura do R-1 e os valores estimados pelo algoritmo de regressão linear na tarefa de sumarização monodocumento.

Por fim, duas análises são realizadas para verificar o impacto dos grupos de características no desempenho do algoritmo de regressão linear com base nas medidas do R-1 e R-2. Primeiro, inicia-se com todas as cem características e remove-se cada um dos quatro grupos de características. Depois, verifica-se o desempenho da regressão linear treinada somente com os atributos de cada um dos grupos de atributos. A Tabela 31 apresenta os resultados deste experimento com base nas medidas do ROUGE.

Remover ou considerar somente um grupo de características, em geral, diminui o de-

Tabela 31 – Resultados (%) da avaliação do impacto dos grupos de atributos no algoritmo de regressão linear com base nas medidas do R-1 e R-2. Os melhores resultados em cada corpus são destacados em negrito, e os cenários de equivalência estatística $p - valor >= 0.05$ são destacados usando o símbolo †.

Cenários de Avaliação	**CNN**	
	R-1	R-2
Todas as Características	**58,58** (19,87)	**41,36** (25,43)
-Ponderação Conceitos	58,53† (19,87)	41,29† (25,49)
-Medidas de Similaridade	58,25 (19,79)	40,95 (25,50)
-Medidas de Divergência	58,24 (19,90)	40,75 (25,63)
-Outras Características	58,35† (19,90)	40,83 (25,63)
Nenhuma Característica		
+Ponderação Conceitos	43,94 (22,56)	29,24 (25,29)
+Medidas de Similaridade	57,96 (19,69)	40,13 (25,35)
+Medidas de Divergência	57,90 (19,84)	40,13 (25,75)
+Outras Características	57,55 (20,12)	40,12 (25,84)

Cenários de Avaliação	**DUC 2001**		**DUC 2002**	
	R-1	R-2	R-1	R-2
Todas as Características	**46,37** (9,76)	**21,10** (11,71)	**49,78** (8,40)	**23,92** (9,74)
-Ponderação Conceitos	46,24† (9,70)	20,98† (11,59)	49,73† (8,41)	**23,92** (9,72)
-Medidas de Similaridade	46,15† (9,46)	20,70† (11,60)	49,44† (8,58)	23,64† (9,79)
-Medidas de Divergência	46,18† (9,79)	20,80† (11,87)	49,52† (8,45)	23,71† (9,77)
-Outras Características	46,07† (9,92)	20,84† (11,83)	49,54† (8,43)	23,69† (9,53)
Nenhuma Característica				
+Ponderação Conceitos	42,69 (10,95)	18,73 (11,33)	44,73 (10,84)	21,05 (10,07)
+Medidas de Similaridade	45,97† (9,67)	20,66† (11,44)	49,52† (8,52)	23,84† (9,88)
+Medidas de Divergência	45,46 (9,78)	20,06 (11,64)	48.98 (8.62)	22.99 (9.95)
+Outras Características	45,18 (9,78)	20,21 (11,29)	48.38 (9.10)	23.02 (9.75)

sempenho do algoritmo de regressão linear em todos os corpora adotados. A única exceção foi no DUC 2002, no qual, remover os atributos relativos ao grupo de *Ponderação dos Conceitos* não impactou no desempenho do algoritmo com base no R-2. Observando somente os cenários de remoção de um grupo de características, é possível observar que não houve uma acentuada perda nos resultados nas medidas do R-1 e R-2. Nos corpora do DUC 2001 e DUC 2002, a depreciação não foi significativa em nenhum cenário. Já no corpus CNN, remover os atributos dos grupos das *Medidas de Similaridade* e *Medidas de Divergência* ocasionou uma redução estatisticamente diferente nas medidas do R-1 e R-2. Ainda nesse corpus, a remoção das características do grupo *Outras Característica* gerou uma diminuição considerável na medida do R-2.

Os cenários considerando somente os atributos de um grupo apresentaram significativas diminuição no desempenho do algoritmo de regressão linear em ambas as medidas do ROUGE. As únicas exceções ocorreram nos corpora do DUC, utilizando as características geradas pelas medidas de similaridades. Os atributos gerados pelas medidas de similaridade e divergência são claramente os mais úteis para o processo de seleção do resumo

mais informativo. Considerar somente esses dois grupo de atributos isoladamente diminui o desempenho do algoritmo, mas essa redução não tão significativa quanto considerar somente os atributos dos grupos de *Outras Características* e *Ponderação Conceitos*.

Os resultados dessas análises demonstram que cada grupo de atributos contribui para o bom desempenho do algoritmo de regressão linear na tarefa de seleção do resumo mais informativo. Os atributos gerados pelas medidas de similaridade e divergência são os mais relevantes para a tarefa, seguidos pelas características dos grupos de *Outras Características* e *Ponderação Conceitos*.

Multidocumento

A quantidade de resumos candidatos produzidos para a tarefa de sumarização multidocumento é muito menor do que na sumarização monodocumento. Por isso, decidiu-se utilizar o método de validação cruzada (*4-fold cross validation*) em nível de corpora, usando um corpus para teste e os outros três corpora para treinamento, como realizado por Hong, Marcus e Nenkova (2015). Com isso, uma maior quantidade de resumos pode ser utilizada para construir o modelo de regressão durante a etapa de treinamento. Para cada corpus, o processo de avaliação é executado da seguinte maneira:

- **Treinamento:** Os resumos candidatos de três corpora são usados nesta etapa para treinar o modelo de regressão.
- **Teste:** O corpus não selecionado durante o treinamento é usado como conjunto de teste. O modelo de regressão construído anteriormente é aplicado para estimar a medida de cobertura do R-1 dos resumos de candidatos das coleções de documentos do corpus de teste. Para cada coleção, seleciona-se o resumo candidato com maior valor estimado do R-1 como o mais informativo.

Na Tabela 32 são apresentados os resultados da avaliação comparativa entre os cinco algoritmos de regressão na tarefa de sumarização multidocumento. No DUC 2001, o algoritmo SMOreg apresentou melhor desempenho nas medidas de correlação de *Pearson* e RSS, enquanto que o algoritmo LeastMedSq obteve maior correlação de *Spearman*. O algoritmo SMOreg obteve melhorias estatísticas nas medidas de cobertura do R-1 e R-2 em relação a todos os demais algoritmos. O SMOreg também conseguiu bons resultados no DUC 2002, sendo superior aos demais algoritmos em todas as medidas de avaliação. Apesar de ter obtido o melhor desempenho nas medidas do ROUGE, essa superioridade só foi estatisticamente significante em relação aos resultados do algoritmo de regressão linear.

No DUC 2003, o algoritmo LeastMedSq apresentou a maior correlação de *Spearman* e menor erro na medida de RSS. Já o algoritmo SMOreg obteve maior correlação de *Pearson*. Em relação às medidas do R-1 e R-2, o algoritmo de regressão RBF apresentou os melhores

Tabela 32 – Resultados da análise comparativa dos cinco algoritmos de regressão para a tarefa de sumarização multidocumento. Em negrito estão destacados os algoritmos com melhor desempenho em cada medida de avaliação e corpus adotado. O símbolo † nas medidas do R-1 e R-2 indicam cenários de equivalência estatística com o algoritmo de melhor desempenho (em negrito).

Algoritmos	**DUC 2001**				
	P	S	RSS (%)	R-1 (%)	R-2 (%)
LeastMedSq	0,6517	0,5490	0,389	34,45 (6,94)	8,32 (5,47)
MLP	0,5951	0,4858	0,388	34,53 (7,00)	8,68 (5,96)
Regressão Linear	0,6499	0,5458	0,386	34,23 (7,19)	8,06 (5,59)
Regressão RBF	0,6444	0,5235	0,376	34,71 (6,83)	8,37 (5,47)
SMOreg	0,6519	0,5377	0,370	**35,74** (7,30)	**9,34** (5,89)
Algoritmos	**DUC 2002**				
	P	S	RSS (%)	R-1 (%)	R-2 (%)
LeastMedSq	0,6517	0,5299	0,304	35,78† (7,26)	8,82† (4,08)
MLP	0,6010	0,4960	0,485	35,60† (7,08)	8,65† (3,66)
Regressão Linear	0,6435	0,5136	0,380	35,47 (6,92)	8,42 (3,86)
Regressão RBF	0,6410	0,5105	0,286	36,48† (5,44)	8,97† (3,87)
SMOreg	0,6791	0,5416	**0,261**	**36,98** (5,66)	**9,20** (3,94)
Algoritmos	**DUC 2003**				
	P	S	RSS (%)	R-1 (%)	R-2 (%)
LeastMedSq	0,6375	0,5428	**0,337**	41,87† (6,11)	11,26† (4,27)
MLP	0,5976	0,5092	0,712	40,62 (5,97)	11,06† (3,89)
Regressão Linear	0,6148	0,5217	0,381	41,41† (6,38)	11,22† (4,47)
Regressão RBF	0,6169	0,4999	0,370	**42,12** (6,03)	**11,68** (4,49)
SMOreg	0,6387	0,5336	0,362	41,39† (6,12)	11,22† (4,51)
Algoritmos	**DUC 2004**				
	P	S	RSS (%)	R-1 (%)	R-2 (%)
LeastMedSq	0,7368	**0,6182**	0,190	39,73† (4,45)	10,37† (3,36)
MLP	0,6830	0,5377	0,375	38,72 (4,67)	9,56 (3,30)
Regressão Linear	0,7230	0,6090	0,198	39,61† (4,38)	10,31† (3,33)
Regressão RBF	0,7291	0,6084	0,219	39,49† (4,71)	10,22† (3,45)
SMOreg	**0,7416**	0,6130	0,184	**40,12** (4,19)	**10,49** (3,35)

resultados. Contudo uma diferença estatística só foi observada quando comparada com os resultados do algoritmo MLP na medida do R-1.

No DUC 2004, o algoritmo SMOreg obteve a melhor performance em todas as medidas de avaliação, com exceção da correlação de *Spearman*, cujo melhor resultado foi obtido pelo algoritmo LeastMedSq. Apesar de ter obtido os melhores resultados nas medidas do ROUGE, essa melhoria só foi significativa em relação ao algoritmo MLP em ambas as medidas do R-1 e R-2.

Assim como na sumarização monodocumento, neste experimento, observou-se a existência de diversos exemplos representando falso positivos, ou seja, resumos com baixo escore no R-1, mas com atributos muito semelhantes com resumos com alto valor na medida do R-1. Tais resumos representam um desafio, já que eles interferem nas etapas de

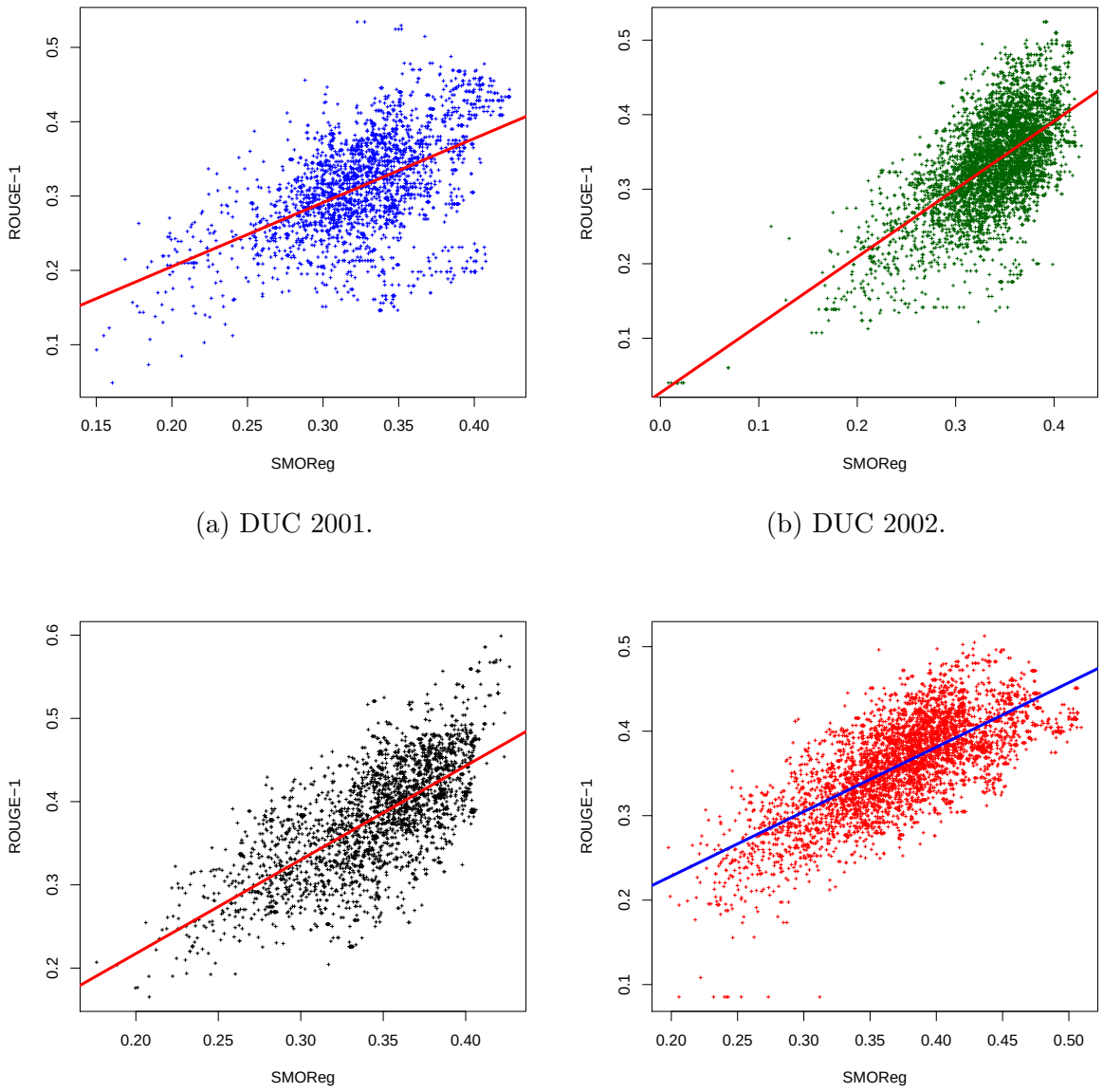
treinamento e teste.

Uma maior diversidade no desempenho dos algoritmos pode ser observada neste experimento para a sumarização multidocumento do que na avaliação do cenário monodocumento. A tarefa de sumarização de múltiplos documentos representa um desafio mais complexo do que a monodocumento, já que a quantidade de informações a ser sintetizada é muito maior. Os algoritmos de regressão obtiveram correlações mais fortes do que usando as características individuais, demonstrando que combinar os atributos produz melhores resultados do que considerá-los isoladamente. O algoritmo SMOreg foi o mais estável e obteve melhores resultados com base nas medidas do R-1 e R-2 do que os demais algoritmos avaliados, na maioria dos cenários de avaliação. Por isso, ele foi selecionado como o melhor algoritmo neste experimento.

A Figura 14 ilustra as correlações de *Pearson* entre os escores do R-1 e os valores estimados pelo algoritmo SMOreg nos quatro corpora do DUC. O SMOreg apresenta uma correlação positiva moderada nos corpora do DUC 2001-2003, e uma forte correlação no DUC 2004. Mesmo assim, é possível observar a presença de diversos resumos, nos quais, o algoritmo estimou valores bem diferentes dos reais escores do R-1.

Na Tabela 33 são apresentados os resultados da avaliação do impacto dos grupos de atributos no desempenho do algoritmo SMOreg em termos das medidas de cobertura do R-1 e R-2. Em geral, a remoção dos grupos de atributos ocasionou uma diminuição no desempenho do algoritmo, mas em três cenários de remoção o desempenho foi melhorado. Remover o grupo de características da *Ponderação dos Conceitos* aumentou os resultados do R-1 nos corpora do DUC 2002 e DUC 2004, enquanto que remover os atributos gerados pelas *Medidas de Similaridade* levou a um melhor desempenho na medida do R-2 no corpus do DUC 2003. Nesses cenários de melhoria, houve um aumento na performance em alguns corpora, enquanto que em outros houve uma diminuição. Com isso, não existe nenhuma garantia de que os atributos que apresentaram um bom desempenho em um corpus, também manterão os bons resultados em um outro corpus. Contudo, usando todas a características é mantido uma estabilidade no desempenho da abordagem proposta em todos os quatro corpora avaliados.

Na maioria dos cenários de avaliação, houve uma diminuição considerável no desempenho do algoritmo SMOreg nas avaliações considerando somente os atributos de cada um dos grupos isoladamente. É possível observar que as características geradas pelas medidas de similaridade e divergência são as mais eficientes para o processo de seleção do resumo mais informativo. Em geral, usando somente os atributos de similaridade ou divergência houve uma perda menor do que considerando somente as características do grupo de ponderação conceitos e outras características. Usando todos os atributos é possível produzir melhores resultados, e também manter um maior equilíbrio no desempenho do algoritmo SMOreg, nos quatro corpora considerados.



(a) DUC 2001.

(b) DUC 2002.

(c) DUC 2003.

(d) DUC 2004.

Figura 14 – Correlação de *Pearson* entre a medida de cobertura do R-1 e os valores estimados pelo algoritmo SMOReg na tarefa de sumarização multidocumento.

6.2.4 Comparação com o Estado da Arte

Este último experimento compara o desempenho da abordagem proposta com outros sistemas do estado da arte nas tarefas de sumarização monodocumento e multidocumento. Para essa comparação, o algoritmo de Regressão Linear (monodocumento) e o SMOreg (multidocumento) são adotados por terem obtido o melhor desempenho, com base nas medidas de avaliação do ROUGE nos experimentos descritos na seção anterior.

Tabela 33 – Resultados (%) da avaliação do impacto dos grupos de atributos no algoritmo SMOreg em termos das medidas do R-1 e R-2. Os melhores resultados em cada corpus são destacados em negrito, e os cenários de equivalência estatística $p - valor > 0.05$ são destacados usando o símbolo †.

Cenários de Avaliação	**DUC 2001**		**DUC 2002**	
	R-1	R-2	R-1	R-2
Todas as Características	**35,74** (7,30)	**9,34** (5,89)	36,98† (5,66)	9,20 (3,94)
-Ponderação Conceitos	35,51† (7,33)	9,29† (6,03)	**37,04** (5,79)	9,27† (4,10)
-Medidas de Similaridade	34,99† (7,38)	8,71† (5,76)	36,63† (5,53)	9,18† (4,01)
-Medidas de Divergência	34,91† (7,33)	8,43† (5,37)	36,60† (5,36)	9,03† (3,79)
-Outras Características	34,38 (7,22)	8,47† (5,61)	36,46† (5,42)	9,04† (3,93)
Nenhuma Característica				
+Ponderação Conceitos	32,18 (7,09)	6,56 (3,78)	35,38 (5,47)	8,94 (3,84)
+Medidas de Similaridade	34,55† (7,23)	8,34† (5,56)	36,42† (5,48)	8,98† (4,00)
+Medidas de Divergência	33,78 (7,18)	8,01 (5,58)	36,86† (5,65)	**9,46†** (4,27)
+Outras Características	33,95 (6,69)	8,35† (4,68)	36,25† (5,52)	8,63† (4,00)
Cenários de Avaliação	**DUC 2003**		**DUC 2004**	
	R-1	R-2	R-1	R-2
Todas as Características	41,39 (6,12)	11,22 (4,51)	40,12 (4,19)	**10,49** (3,35)
-Ponderação Conceitos	40,99† (5,64)	11,04† (4,40)	**40,30†** (4,22)	10,46† (3,15)
-Medidas de Similaridade	**42,01†** (6,32)	11,09† (4,63)	39,62† (4,72)	10,19† (3,34)
-Medidas de Divergência	41,35† (5,67)	11,27† (4,30)	39,81† (4,52)	10,33† (3,48)
-Outras Características	40,55† (5,96)	11,13† (4,60)	39,07 (4,20)	10,32† (3,26)
Nenhuma Característica				
+Ponderação Conceitos	39,47 (6,78)	10,52† (4,96)	38,03 (4,41)	9,79 (3,15)
+Medidas de Similaridade	40,98† (6,04)	**11,58** (4,52)	39,25† (4,40)	10,11† (3,21)
+Medidas de Divergência	40,96† (6,15)	11,49† (4,79)	39,53† (4,17)	10,14† (3,02)
+Outras Características	37,83 (6,57)	9,53 (4,47)	38,25 (4,35)	10,10† (3,31)

Monodocumento

Este experimento, na tarefa de sumarização monodocumento, compara a performance da abordagem proposta neste capítulo com os seguintes sistemas do estado da arte:

1. O *baseline* consistindo do tradicional método de posição das sentenças que seleciona as n primeiras sentenças do documento de entrada. No corpus CNN, n depende do número de sentenças do documento, enquanto que nos corpora do DUC, seleciona-se as n primeiras sentenças até que o limite de 100 palavras seja alcançado;

2. Os sistemas T e 28 foram os melhores participantes da competição original do DUC 2001-2002 identificados nos experimentos conduzidos neste trabalho, respectivamente;

3. Os sistemas AutoSummarizer, Classifier4J, e HP-UFPE FS. Esses sistemas apresentaram os melhores resultados nos experimentos descritos em Batista et al. (2015);

4. O sistema baseado em conceitos usando PLI proposto no Capítulo 4; e

5. Os *Oráculos* que representam sistemas hipotéticos que sempre selecionam corretamente o resumo com o maior valor do R-1 entre o conjunto de resumos de candidatos para cada documento de entrada.

A Tabela 34 apresenta os resultados obtidos em termos das medidas do R-1 e R-2 da comparação entre a abordagem proposta usando o algoritmo de regressão linear e os sistemas da literatura mencionados anteriormente. No corpus CNN, a abordagem proposta obteve significativa melhoria de performance com base na medida do R-1 em relação a todos os demais sistemas. Apesar de ter apresentado também a maior média de cobertura no R-2, essa superioridade não foi estatística comparada com a abordagem baseada em conceitos proposta no Capítulo 4 e usada aqui para gerar os resumos candidatos.

Tabela 34 – Resultados comparativos (%) e desvio padrão entre parênteses em relação a outras abordagens do estado da arte para sumarização monodocumento. O melhor desempenho global de cada corpus é destacado em negrito, e o grupo de sistemas estatisticamente semelhantes, se existir, é indicado por †.

Sistemas	**CNN**	
	R-1	R-2
AutoSummarizer	48,81 (18,70)	32,74 (22,70)
Baseline	45,99 (21,77)	33,49 (25,00)
Classifier4J	46,63 (20,32)	32,15 (23,13)
Abordagem PLI	57,54 (20,09)	41,08† (25,38)
HP-UFPE FS	50,71 (20,34)	34,58 (24,38)
Proposta	**58,58** (19,87)	**41,36** (25,43)
Oráculo	*74,13* (16,96)	*65,71* (19,00)

Sistemas	**DUC 2001**		**DUC 2002**	
	R-1	R-2	R-1	R-2
AutoSummarizer	41,92 (9,04)	16,63 (9,95)	43,79 (8,78)	19,17 (9,31)
Baseline	43,75 (10,47)	19,57 (11,64)	46,94 (9,20)	22,14 (10,01)
Classifier4J	44,44 (9,85)	19,86 (11,34)	47,09 (8,93)	22,12 (9,87)
Abordagem PLI	45,32 (9,74)	20,25 (11,52)	48,85 (8,45)	23,30 (9,76)
HP-UFPE FS	35,91 (11,78)	11,78 (9,78)	45,70 (9,31)	20,59 (9,88)
Sistema T/28	44,53 (9,23)	20,27† (10,75)	48,07 (8,90)	22,88 (9,96)
Proposta	**46,37** (9,76)	**21,10** (11,71)	**49,78** (8,40)	**23,92** (9,74)
Oráculo	*53,33* (8,81)	*28,02* (12,36)	*56,16* (7,87)	*30,39* (10,50)

A abordagem proposta também apresentou melhorias significativas nos corpora do DUC 2001-2002, especialmente considerando à medida do R-1. Em relação à medida do R-2, ela também obteve a melhor performance, mas os resultados não foram tão bons quanto os do R-1. Mesmo assim, somente no DUC 2001, o sistema T apresentou desempenho estatisticamente semelhante considerando a medida do R-2.

Em geral, a abordagem proposta demonstrou ter uma eficiente performance, apresentando melhorias significativas em comparação com outros sistemas do estado da arte. Na

Tabela 35 e Tabela 36 são apresentados os p-valores obtidos aplicando os teste de *Wilcoxon signed rank* no corpora do CNN e DUC, respectivamente. Os p-valores indicando uma diferença estatisticamente significante ao nível de confiança de 95% ou mais são destacados em negrito. É possível observar que a abordagem proposta é superior aos demais sistemas com um nível de significância menor do 0.01 em todos os cenários de avaliação.

Tabela 35 – P-valores obtidos aplicando o teste de *Wilcoxon signed rank* no corpus CNN.

CNN					
Sistemas	AutoS	Baseline	C4J	HP-UFPE	Proposta
Ab. PLI	(>) **1.35e-134**	(>) **8.88e-188**	(>) **2.83e-201**	(>) **7.37e-99**	(<) **5.89e-09**
AutoS	-	(>) **2.04e-11**	(>) **7.19e-09**	(<) **1.31e-10**	(<) **3.27e-167**
Baseline	-	-	(<) 0.21	(<) **2.88e-44**	(<) **4.68e-198**
C4J	-	-	-	(<) **1.15e-37**	(<) **1.09e-215**
HP-UFPE	-	-	-	-	(<) **2.47e-128**

Tabela 36 – P-valores obtidos aplicando o teste de *Wilcoxon signed rank* nos corpora do DUC 2001-2002.

DUC 2001						
Sistemas	AutoS	Baseline	C4J	HP-UFPE	Sist. T	Proposta
Ab. PLI	(>) **1.46e-10**	(>) **9.38e-05**	(>) **0.014**	(>) **1.73e-32**	(>) **0.019**	(<) **3.55e-5**
AutoS.	-	(<) **1.41e-03**	(<) **9.20e-06**	(>) **1.16e-16**	(<) **5.31e-04**	(<) **1.04e-17**
Baseline	-	-	(<) **7.57e-04**	(>) **2.12e-24**	(<) 0.62	(<) **2.83e-10**
C4J	-	-	-	(>) **1.22e-28**	(<) 0.50	(<) **1.68e-07**
HP-UFPE	-	-	-	-	(<) **1.43e-19**	(<) **1.72e-39**
Sist. T	-	-	-	-	-	(<) **0.0001**
DUC 2002						
Sistemas	AutoS	Baseline	C4J	HP-UFPE	Sist. 28	Proposta
Ab. PLI	(>) **1.70e-29**	(>) **5.09e-13**	(>) **1.01e-10**	(>) **2.22e-19**	(>) **0.001**	(<) **3.94e-05**
AutoS	-	(<) **3.17e-10**	(<) **4.86e-12**	(<) **1.60e-04**	(<) **4.23e-22**	(<) **1.90e-44**
Baseline	-	-	(<) 0.469	(>) **2.71e-04**	(<) **7.35e-06**	(<) **4.21e-22**
C4J	-	-	-	(>) **2.37e-05**	(<) **2.16e-04**	(<) **2.74e-19**
HP-UFPE	-	-	-	-	(<) **4.52e-12**	(<) **1.20e-29**
Sist. 28	-	-	-	-	-	(<) **2.04e-10**

Apesar dos bons resultados, ainda há muito espaço para melhorias. Os resultados obtidos pela abordagem proposta ainda estão muito distantes dos valores obtidos pelo sistema Oráculo nos três corpora. Isso demonstra que a etapa de geração dos resumos candidatos foi capaz de produzir resumos informativos para os documentos de entrada, mas esses não foram selecionados na etapa de seleção do resumo mais informativo. Esse problema ocorre por erros cometidos pelos algoritmos de regressão durante a estimação da medida do R-1 e consequentemente na seleção do resumo mais informativo.

Multidocumento

Neste experimento, avalia-se o desempenho da abordagem proposta usando o algoritmo SMOreg na tarefa de sumarização multidocumento em relação aos seguintes sis-

temas do estado da arte: **(i)** os sistemas participantes nas competições DUC 2001-2004 com melhor desempenho encontrados nos experimentos realizados neste trabalho; **(ii)** os seguintes sistemas do estado da arte: ICSISumm (GILLICK et al., 2009), Greedy-KL (HAGHIGHI; VANDERWENDE, 2009), LLRSum (CONROY; SCHLESINGER; O'LEARY, 2006), ProbSum (NENKOVA; VANDERWENDE; MCKEOWN, 2006) e Sume (BOUDIN; MOUGARD; FAVRE, 2015); e **(iii)** a abordagem baseada em conceitos usando PLI para a sumarização multidocumento proposta no Capítulo 5.

Os resultados deste experimento com base nas medidas do R-1 e R-2 são apresentados na Tabela 37 e os p-valores obtidos pelo teste de *Wilcoxon signed rank* são demonstrados na Tabela 39. Cenários de diferença estatística ao nível de confiança de 95% ou mais são destacados em negrito. A abordagem proposta apresentou a melhor performance com base em ambas as medidas do ROUGE em três dos quatro corpora, enquanto que o sistema ICSISumm obteve melhor desempenho no DUC 2002.

Tabela 37 – Resultados (%) e desvio padrão (entre parênteses) das comparações entre a abordagem proposta e outros sistemas em termos das medidas do R-1 e R-2. O sistema de melhor desempenho em cada corpus é destacado em negrito. Cenários de diferença estatística entre o sistema de melhor desempenho e os outros sistemas são indicados usando o símbolo †.

Sistemas	**DUC 2001**		**DUC 2002**	
	R-1	R-2	R-1	R-2
Abordagem PLI	34,48† (6,90)	8,51† (4,24)	36,48† (5,65)	9,26† (3,76)
Greedy-KL	32,84 (6,43)	6,70 (3,64)	35,79 (5,74)	7,49 (3,61)
ICSISumm	33,88 (6,95)	7,75 (4,30)	**37,34** (5,05)	**9,53** (3,83)
LLRSum	32,00 (5,88)	6,76 (3,25)	32,84 (5,55)	6,75 (3,72)
Proposta	**35,74** (7,30)	**9,34** (5,89)	36,98† (5,66)	9,20† (3,94)
ProbSum	29,73 (5,41)	5,16 (2,64)	32,57 (4,74)	7,06 (3,63)
Sume	33,37 (7,14)	7,73 (4,29)	34,30 (5,26)	8,15 (3,92)
Sistema P/26	31,69 (6,43)	6,30 (3,76)	35,21 (5,30)	7,66 (3,30)
Oráculo	*39,58* (6.65)	*11,71* (5,78)	*42,20* (4,57)	*11,37* (4,53)
Sistemas	**DUC 2003**		**DUC 2004**	
	R-1	R-2	R-1	R-2
Abordagem PLI	40,47† (6,00)	10,95† (4,12)	39,05 (4,42)	10,04 (3,25)
Greedy-KL	40,35† (5,76)	9,20 (3,99)	38,27 (4,73)	8,96 (3,09)
ICSISumm	40,07† (4,88)	10,95† (4,00)	38,42 (4,14)	9,80 (3,17)
LLRSum	36,94 (6,05)	8,87 (3,21)	35,90 (5,01)	8,06 (3,12)
ProbSum	37,60 (7,11)	9,28 (4,05)	35,37 (4,41)	8,18 (3,00)
Proposta	**41,39** (6,12)	**11,22** (4,51)	**40,12** (4,19)	**10,49** (3,35)
Sume	39,36† (5,57)	9,81† (3,95)	37,29 (4,24)	8,83 (2,71)
Sist. 12/Classy 04	38,44 (5,25)	9,11 (3,95)	37,69 (4,08)	8,98 (3,08)
Oráculo	*46,34* (5,22)	*14,01* (4,84)	*43,32* (3,73)	*11,77* (3,07)

A solução proposta via PLI e regressão apresentou melhorias estatísticas em relação aos demais sistemas considerados nos corpora do DUC 2001 e DUC 2004 nas duas medidas

do R-1 e R-2, com exceção dos resultados obtidos pela abordagem proposta no Capítulo 5 usando somente PLI. Apesar de ter atingido a melhor performance em termos do R-1 e R-2 no corpus do DUC 2003, seu desempenho pode ser considerado estatisticamente similar aos sistemas ICSISumm, Greedy-KL, Sume e ao sistema proposto usando somente PLI. No corpus do DUC 2002, o sistema ICSISumm obteve os melhores resultados, mas as duas abordagens propostas neste trabalho apresentam desempenho estatisticamente similar nas duas medidas do ROUGE. A abordagem proposta neste capítulo usando PLI e regressão melhorou os resultados do sistema baseado em conceitos adotado para gerar os resumos candidatos, mas ainda assim não obteve melhorias em relação ao ICSISumm.

Assim como aconteceu na sumarização monodocumento, o desempenho da abordagem proposta ainda está muito distante dos resultados obtidos pelos sistemas de Oráculo, que representam o limite superior que pode ser obtido selecionando sempre o resumo candidato com maior valor real do R-1. Esses resultados demonstram que o processo de geração dos resumos candidatos consegue produzir resumos informativos, ou seja, resumos com alto escore na medida do R-1, mas a etapa de seleção do resumo mais informativo não conseguiu identificar corretamente o candidato mais informativo. Isso acontece por erros durante a estimação da medida do R-1 pelo algoritmo de regressão, ocasionando a seleção de um candidato com valor do R-1 menor do que outros candidatos disponíveis.

Como mencionado anteriormente, o sistema SumCombine proposto por Hong, Marcus e Nenkova (2015) é similar a abordagem proposta neste capítulo, no sentido que gera diversos resumos candidatos, e posteriormente, aplica um algoritmo de regressão para estimar e selecionar o resumo mais informativo. Infelizmente, nem a implementação do SumCombine nem os resumos gerados estão disponíveis. Dessa forma, não é possível compará-lo com a abordagem proposta de forma direta, pois os resumos gerados são avaliados em ambientes experimentais diferentes. Por isso, realizou-se uma comparação indireta, avaliando os ganhos obtidos pelo SumCombine e pela abordagem proposta em relação ao sistema ICSISumm. A Tabela 38 resume os resultados obtidos neste experimento com base nas medidas do R-1 e R-2.

Analisando a Tabela 38 é possível observar que a abordagem proposta neste capítulo apresentou um ganho médio maior nos corpora do DUC 2001 e DUC 2003, enquanto que o SumCombine obteve maiores ganhos nos corpora do DUC 2002 e DUC 2004. Ambos os sistemas apresentam ganhos médios similares nos corpora do DUC 2001, DUC 2003 e DUC 2004. Contudo, uma diferença mais acentuada pode ser vista no DUC 2002. O SumCombine por utilizar os resumos gerados pelo sistema ICSISumm, consegue apresentar ganhos significativos neste corpus. A abordagem proposta adota o método baseado em conceitos apresentado no Capítulo 5 para gerar os resumos candidatos. Como esse método obteve um desempenho muito inferior ao sistema ICSISumm neste corpus, isso acabou refletindo também na abordagem via regressão proposta neste capítulo.

De maneira global, levando em consideração os ganhos médios nos quatro corpora in-

Tabela 38 – Resultados (%) dos ganhos médios obtidos nas medidas de cobertura do R-1 e R-2 pelo sistema SumCombine e pela abordagem proposta, em relação ao sistema ICSISumm.

Hong et al. (2015)	**DUC 2001**		**DUC 2002**		**DUC 2003**		**DUC 2004**	
	R-1	R-2	R-1	R-2	R-1	R-2	R-1	R-2
ICSISumm	34,21	7,85	37,35	9,53	38,13	10,28	38,41	9,78
SumCombine	35,26	7,88	38,23	9,46	39,59	10,18	39,95	10,48
Ganho Médio	1,05	0,03	**0,88**	**-0,07**	**1,46**	-0,1	1,54	**0,7**
Abordagem Proposta	**DUC 2001**		**DUC 2002**		**DUC 2003**		**DUC 2004**	
	R-1	R-2	R-1	R-2	R-1	R-2	R-1	R-2
ICSISumm	33,88	7,75	37,34	9,53	40,07	10,95	38,42	9,8
Abordagem Proposta	35,74	9,34	36,98	9,2	41,39	11,22	40,12	10,49
Ganho Médio	**1,86**	**1,59**	-0,36	-0,33	1,32	**0,27**	**1,7**	0,69

Tabela 39 – P-valores obtidos aplicando o teste de *Wilcoxon signed rank* para comparar os sistemas do estado da arte na tarefa de sumarização multidocumento.

DUC 2001							
Sistemas	Greedy KL	ICSISumm	LLRSum	Sist, P	ProbSum	Proposta	Sume
Ab. PLI	(>) 0,092	(>) 0,894	(>) **0,010**	(>) **0,007**	(>) **5,74e-05**	(<) 0,060	(>) 0,106
Greedy KL	-	(<) 0,149	(>) 0,593	(>) 0,290	(>) **1,95e-03**	(<) **0,002**	(<) 0,322
ICSISumm	-	-	(>) 0,058	(>) **0,012**	(>) **1,05e-04**	(<) 0,050	(>) 0,546
LLRSum	-	-	-	(>) 0,636	(>) **0,006**	(<) **0,001**	(<) 0,190
Sist. P	-	-	-	-	(>) **0,015**	(<) **0,0001**	(<) 0,166
ProbSum	-	-	-	-	-	(<) **1,95e-05**	(<) **9,58e-04**
Proposta	-	-	-	-	-	-	(>) **0,003**
DUC 2002							
Sistemas	Greedy KL	ICSISumm	LLRSum	Sist, 26	ProbSum	Proposta	Sume
Ab. PLI	(>) 0,408	(<) 0,155	(>) **3,17e-06**	(>) 0,196	(>) **4,22e-06**	(<) 0,140	(>) **1,66e-04**
Greedy KL	-	(<) **0,012**	(>) **1,50e-04**	(>) 0,425	(>) **4,48e-04**	(<) 0,054	(>) 0,054
ICSISumm	-	-	(>) **1,19e-06**	(>) **0,0007**	(>) **4,78e-08**	(>) 0,492	(>) **1,44e-06**
LLRSum	-	-	-	(<) 0,636	(>) **0,006**	(<) **0,001**	(<) 0,190
Sist. 26	-	-	-	-	(>) **0,002**	(<) **0,016**	(>) 0,239
ProbSum	-	-	-	-	-	(<) **7,33e-07**	(<) 0,050
Proposta	-	-	-	-	-	-	(>) **1,59e-05**

DUC 2003							
Sistemas	Greedy KL	ICSISumm	LLRSum	Sist, 12	ProbSum	Proposta	Sume
Ab. PLI	(>) 0,641	(>) 0,364	(>) **1,56e-04**	(>) **0,004**	(>) **0,008**	(<) 0,629	(>) 0,161
Greedy KL	-	(>) 0,715	(>) **7,92e-04**	(>) **0,022**	(>) **0,001**	(<) 0,299	(>) 0,309
ICSISumm	-	-	(>) **0,001**	(>) 0,088	(>) 0,063	(<) 0,150	(>) 0,561
LLRSum	-	-	-	(<) 0,099	(<) 0,424	(<) **3,92e-05**	(<) **0,003**
Sist. 12	-	-	-	-	(>) **0,002**	(<) **0,016**	(>) 0,239
ProbSum	-	-	-	-	-	(<) **0,0008**	(<) 0,088
Proposta	-	-	-	-	-	-	(>) 0,094
DUC 2004							
Sistemas	Greedy KL	ICSISumm	LLRSum	Classy 04	ProbSum	Proposta	Sume
Ab. PLI	(>) 0,086	(>) 0,203	(>) **1,12e-06**	(>) 0,015	(>) **9,92e-06**	(<) **0,003**	(>) **0,003**
Greedy KL	-	(<) 0,528	(>) **5,87e-05**	(>) 0,292	(>) **1,69e-04**	(<) **0,0003**	(>) 0,141
ICSISumm	-	-	(>) **4,10e-05**	(>) 0,114	(>) **5,64e-05**	(<) **0,0009**	(>) **0,017**
LLRSum	-	-	-	(<) **0,015**	(>) 0,301	(<) **2,31e-08**	(<) **0,005**
Classy 04	-	-	-	-	(>) **0,002**	(<) **0,016**	(>) 0,239
ProbSum	-	-	-	-	-	(<) **1,22e-07**	(<) **0,001**
Proposta	-	-	-	-	-	-	(>) **4,56e-05**

vestigados, o SumCombine apresentou um ganho médio em relação ao sistema ICSISumm de 1,23% na medida do R-1 e 0,14% no R-2. Já a abordagem proposta neste capítulo apresentou um ganho médio de 1,13% no R-1 e 0,56% no R-2 comparado com o ICSISumm. Esses resultados indicam que ambas as abordagens apresentam um desempenho similar nas medidas do R-1 e R-2, com a abordagem proposta obtendo um desempenho melhor na medida do R-2. A performance da abordagem proposta é encorajadora quando comparada com os resultados obtidos pelo sistema SumCombine. Vale salientar que o sistema SumCombine utiliza 360 características para construir o seu modelo regressão, além de gerar uma enorme quantidade de resumos candidatos para cada coleção de documentos de entrada: 7,498 (DUC 2001), 12,048 (DUC 2002), 3,448 (DUC 2003) e 3,270 (DUC 2004). Enquanto isso, a abordagem proposta adota somente 100 características no modelo de regressão e produz 100 resumos candidatos para cada coleção de documentos de entrada.

Dado os resultados obtidos e os números mencionados acima, é possível concluir que a abordagem proposta conseguiu obter um desempenho comparável com o sistema SumCombine, utilizando menos atributos para gerar o modelo de regressão e produzindo uma quantidade muito menor de resumos candidatos para cada grupo de documentos de entrada a ser sintetizado.

6.3 Considerações Finais do Capítulo

Este capítulo apresentou uma abordagem para a sumarização automática de textos baseada em conceitos via PLI e regressão. Inicialmente, a abordagem proposta explora um método de sumarização baseado em conceitos com PLI para gerar um conjunto de resumos candidatos. Posteriormente, um modelo de regressão é aplicado para estimar a informatividade de cada candidato gerado, adotando a tradicional medida de cobertura do R-1 como atributo alvo. O modelo proposto adota diversas características identificadas na literatura e novos atributos propostos neste trabalho para medir a cobertura de informações relevantes presentes no resumo, sob diferentes perspectivas.

Os resultados experimentais demonstram que **(i)** os novos atributos propostos com base nos métodos de ponderação dos conceitos de frequência das sentenças (monodocumento) ou dos documentos (multidocumento), posição das sentenças, e o método do ponderação combinado, alcançaram uma correlação mais forte do que as características anteriores adotadas na literatura com base na probabilidade de ocorrência dos n-gramas; e **(ii)** a abordagem proposta é consistente e, em muitos cenários, estatisticamente superior do que muitos dos sistemas de sumarização do estado da arte em vários corpora, com base nas medidas do R-1 e R-2, em ambas as tarefas de sumarização monodocumento e multidocumento.

No próximo capítulo são apresentadas as conclusões do trabalho desenvolvido, suas limitações, as contribuições obtidas e algumas perspectivas de trabalhos futuros.

7 CONCLUSÃO

Neste trabalho de doutorado foi desenvolvida uma abordagem baseada em conceitos utilizando PLI e regressão para as tarefas de sumarização monodocumento e multidocumento de artigos de notícias escritos em Inglês. A arquitetura da abordagem proposta é composta por duas etapas centrais: **(i)** A geração de diversos resumos candidatos; e posteriormente, a **(ii)** Seleção do resumo mais informativo. A etapa de geração dos resumos candidatos é composta por duas abordagens baseadas em conceitos usando PLI (também propostas neste trabalho) que são executadas, dependendo da tarefa em questão, adotando diferentes configurações, por exemplo, formas de representação e métodos de ponderação de conceitos diversos. Após essa etapa de geração, cada resumo candidato é analisado, visando quantificar a sua informatividade. Para isso, um algoritmo de regressão é aplicado para estimar a medida de cobertura do ROUGE-1 em cada candidato gerado, permitindo assim, selecionar o resumo estimado como mais informativo. O modelo de regressão proposto foi construído usando indicadores de relevância de conteúdo, como posição, frequência e centralidade, individualmente e combinados, em conjunto com diversas medidas de similaridade e divergência computados entre o resumo e o(s) documento(s) de entrada.

As abordagens propostas foram avaliadas adotando os principais corpora das áreas de sumarização monodocumento e multidocumento de artigos de notícias escritos em Inglês. Três conjuntos de dados foram considerados para a tarefa de sumarização monodocumento (CNN, DUC 2001 e DUC 2002), enquanto que para a sumarização multidocumento foram adotados os quatro corpora do DUC 2001-2004. Para avaliar os resumos gerados adotou-se, como principal medida de avaliação, os escores de cobertura do ROUGE-1 e ROUGE-2 em todos os experimentos realizados. Diversos aspectos de cada uma das abordagens desenvolvidas foram avaliados e comparados com outros sistemas do estado da arte em ambas as tarefas de sumarização.

Os resultados experimentais obtidos demonstraram que as abordagens baseadas em conceitos usando PLI propostas apresentam desempenho competitivo com os sistemas do estado da arte considerados, nas tarefas de sumarização monodocumento e multidocumento. A estratégia de integração das abordagens baseadas em conceitos combinando PLI e regressão, em uma única macro abordagem, resultou na geração de resumos ainda mais informativos. A macro solução desenvolvida obteve resultados consistentes e estatisticamente superiores em comparação com outros trabalhos da literatura na maioria dos cenários avaliados com base nas medidas do ROUGE adotadas.

7.1 Principais Contribuições e Descobertas

No Capítulo 3, investigou-se o desempenho de diversas técnicas de pontuação de sentenças e estratégias de combinação para as tarefas de sumarização monodocumento e multidocumento no domínio de artigos de notícias. As análises realizadas permitiram observar que as técnicas relacionadas com a posição das sentenças e a frequência das palavras, além de combinações lineares dessas, apresentaram um bom desempenho na sumarização monodocumento com base nas medidas de cobertura do ROUGE-1 e ROUGE-2. Os experimentos no contexto da sumarização multidocumento evidenciaram que os aspectos de posição das sentenças e centralidade das suas informações, além de suas combinações lineares, resultaram, em geral, na geração de resumos informativos com base nas medidas de avaliação do ROUGE. Por fim, os resultados experimentais obtidos demonstraram que nenhuma das técnicas e combinações investigadas conseguiu gerar resumos informativos para todos os documentos de entrada, em ambas as tarefas de sumarização consideradas. Foi inspirada nessa observação que a abordagem baseada em conceitos usando PLI e regressão apresentada neste trabalho foi desenvolvida.

No Capítulo 4, apresentou-se a abordagem baseada em conceitos usando PLI proposta neste trabalho para a sumarização monodocumento. A abordagem criada adota bigramas como conceitos devido aos bons resultados obtidos usando essa forma de representação nos experimentos realizados comparando diversas formas de representação. A adoção do modelo de Grafo de Entidades (GUINAUDEAU; STRUBE, 2013) para aproximar uma pontuação da coesão local do resumo gerado melhorou o desempenho da abordagem proposta, em termos das medidas de cobertura do ROUGE-1 e ROUGE-2, na maioria dos cenários avaliados. Esse bom desempenho demonstrou que selecionar sentenças conectadas que maximizam a cobertura de bigramas relevantes produzem resumos mais informativos do que escolher frases desconectadas. Os resultados obtidos, avaliando diferentes métodos de ponderação de conceitos, confirmam a hipótese de que combinar os métodos de posição das sentenças e frequência das sentenças para mensurar a relevância dos conceitos extraídos produz resumos informativos para a tarefa de sumarização monodocumento.

A estratégia proposta de pontuar somente a primeira ocorrência dos conceitos extraídos gerou melhores resumos com base nas medidas do ROUGE do que a tradicional tática de pontuar todas as ocorrências, em todos os cenários avaliados. Esses resultados validam outra hipótese criada de que resumos informativos podem ser criados selecionando sentenças no início do documento caso eles possuam conceitos relevantes, ou sentenças ao longo do texto caso elas incluam novos conceitos importantes ainda não inseridos no resumo em criação. Por fim, demonstrou-se que a inserção das restrições de legibilidade investigadas, apesar de evitarem problemas de correferência em aberto e quebras nas relações de discurso, também ocasionam uma queda, na maioria dos casos não significativa, no desempenho da abordagem proposta com base nas medidas de avaliação do ROUGE.

No Capítulo 5, foi descrita a abordagem baseada em conceitos usando PLI proposta

para a tarefa de sumarização multidocumento neste trabalho. Os resultados das avaliações de diversas formas de representação de conceitos demonstraram que adotar bigramas produziu melhores resultados com base nas medidas do ROUGE. As avaliações de diferentes métodos de ponderação confirmaram a hipótese de que combinar os métodos de posição das sentenças e frequência dos documentos para mensurar a relevância dos conceitos, na tarefa de sumarização multidocumento, produz resumos mais informativos com base nas medidas de cobertura do ROUGE do que adotar as técnicas individuais. A estratégia de filtragem criada, baseada no percentual de documentos de entrada, que remove grupos de sentenças com poucos membros apresentou bons resultados nos experimentos realizados, diminuindo o tempo de execução do modelo de PLI e também aumentando a informatividade dos resumos gerados com base nas medidas do ROUGE. Esse bom desempenho valida a hipótese de que filtrar sentenças com um baixo grau de centralidade melhora o desempenho das abordagens baseadas em conceitos utilizando PLI, em termos de tempo de execução e também a informatividade dos resumos produzidos.

No Capítulo 6, foi apresentada a abordagem baseada em conceitos proposta utilizando PLI e regressão para as tarefas de sumarização monodocumento e multidocumento. A arquitetura da solução desenvolvida inicia seu processo de sumarização gerando diversos resumos candidatos para cada documento ou coleção de documentos de entrada, adotando as abordagens baseadas em conceitos com PLI propostas no Capítulo 4 e Capítulo 5, dependendo da tarefa de sumarização a ser realizada. Tal estratégia, explora diferentes métodos de ponderação e formas de representação para os conceitos, além de outras configurações específicas de cada abordagem, permitiu a geração de uma grande diversidade de resumos candidatos informativos, conduzindo a elevados limites superiores com base nas medidas de cobertura do ROUGE-1 e ROUGE-2.

Posteriormente, um algoritmo de regressão é aplicado para estimar a medida de cobertura do ROUGE-1, adotada neste trabalho como escore de informatividade, de cada resumo candidato. Dessa forma, é possível selecionar o resumo estimado como mais informativo. Um modelo de regressão foi desenvolvido utilizando como características, diversos indicadores de relevância de conteúdo, como frequência, posição e centralidade, em conjunto com medidas que mensuram a similaridade e divergência entre o resumo analisado e o(s) documento(s) de entrada. Os resultados experimentais obtidos demonstraram que resumos informativos, em ambas as tarefas de sumarização, tendem a possuir uma alta similaridade e uma baixa divergência com o(s) documento(s) de entrada, principalmente adotando os atributos gerados aplicando os métodos de posição das sentenças e o método de ponderação de conceitos combinado proposto. Além disso, resumos informativos também apresentaram uma alta similaridade com outros resumos candidatos gerados, adotando outras configurações para o mesmo documento ou coleção de documentos de entrada. Os novos atributos propostos com base nos métodos de ponderação de conceitos de frequência das sentenças (monodocumento) ou dos documentos (multidocumento),

posição das sentenças, e o método de ponderação combinado apresentaram uma correlação mais forte do que as características anteriores adotadas na literatura com base na probabilidade de ocorrência dos n-gramas.

No Apêndice A, investigou-se os problemas de representar e estimar a importância dos conceitos em uma abordagem usando PLI para a sumarização monodocumento. Tal problema é menos investigado do que outras tarefas, como por exemplo, estimar a relevância das sentenças para compor o resumo a ser gerado. Cinco formas de representação de conceitos foram analisadas, sendo elas, unigramas, bigramas, entidades nomeadas, dependências sintáticas rotuladas e com um rótulo genérico. Os resultados obtidos demonstraram que adotar bigramas como conceitos levou a geração de resumos mais informativos, com base nas medidas de cobertura do ROUGE-1 e ROUGE-2, na maioria dos cenários de avaliação. Com relação às técnicas de ponderação de conceitos analisadas, o método de posição das sentenças obteve resultados significativamente melhores do que todos os outros métodos investigados. O método de frequência das sentenças também obteve bons resultados, apresentando a segunda melhor performance entre os métodos avaliados. Por fim, observou-se que os métodos de posição e frequência das sentenças, além de obterem bons resultados, também geravam resumos com uma alta diversidade. Esse fato inspirou a proposta de combiná-los em um único método de ponderação, conforme apresentado no Capítulo 4.

7.2 Produção Bibliográfica

Nas subseções a seguir são listados os artigos publicados em periódicos e conferências, respectivamente, durante o desenvolvimento deste trabalho de doutorado.

7.2.1 Artigos em Periódicos

1. **Oliveira, H. T. A.**; Mello, R. F. L.; Lima, R. J.; Lins, R. D.; Freitas, F.; Riss, M.; Simske, S. J. Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization. Expert Systems with Applications. v.65, p.68-86, 2016.

7.2.2 Artigos em Conferências

1. **Oliveira, H. T. A.**; Lima, R. J.; Lins, R. D.; Freitas, F.; Riss, M.; Simske, S. J. Assessing Concept Weighting in Integer Linear Programming based Single-document Summarization In: ACM International Symposium on Document Engineering, 2016, Viena.

2. **Oliveira, H. T. A.**; Lima, R. J.; Lins, R. D.; Freitas, F.; Riss, M.; Simske, S. J. A Concept-based Integer Linear Programming Approach for Single-Document

Summarization In: Brazilian Conference on Intelligent Systems - BRACIS, 2016, Recife.

3. **Oliveira, H. T. A.**; Lins, R. D.; Lima, R. J.; Freitas, F.; Simske, S. J. A Regression-based Approach using Integer Linear Programming for Single-document Summarization In: 29th IEEE International Conference on Tools with Artificial Intelligence, 2017, Boston.

7.3 Limitações

A principal limitação das abordagens propostas neste trabalho é a sua natureza extrativa, ou seja, os resumos criados são compostos somente pelas sentenças originais do(s) documento(s) de entrada, sem nenhuma alteração. Resumos criados de forma extrativa, em geral, apresentam problemas de coesão, principalmente relacionados às quebras no fluxo de ideias entre as sentenças e correferências em aberto (CHRISTENSEN et al., 2013). Além disso, os resumos extrativos também são limitados em relação à sua informatividade. As sentenças de um documento tendem a conter fragmentos de informações relevantes e não relevantes ao mesmo tempo. Dessa forma, incluí-las sem nenhuma alteração resulta no desperdício de espaço, que poderia ser usado para inserir outras informações mais importantes, tornando o resumo gerado ainda mais informativo.

Outra limitação das abordagens criadas é o seu viés com o tipo dos documentos de entrada a serem sintetizados. Este trabalho teve como escopo, somente a sumarização de artigos de notícias. Por isso, visando maximizar o seu desempenho, características específicas, identificadas na literatura para esse tipo de documento foram exploradas. Por exemplo, o aspecto de posição das sentenças, que é amplamente explorado neste trabalho, pode não ter um impacto positivo em outros tipos de documentos, por exemplo, em artigos científicos, e-mails, blogs, entre outros. Sendo assim, não existe nenhuma garantia de que as abordagens propostas neste trabalho também apresentem um bom desempenho para outros tipos de documentos textuais, como os mencionados anteriormente.

A adoção das medidas de avaliação do ROUGE também representa uma importante limitação. As medidas do ROUGE, apesar de possuírem uma boa correlação com avaliações humanas (LIN, 2004; OWCZARZAK et al., 2012), são puramente léxicas, ou seja, é realizado o casamento léxico entre os n-gramas do resumo gerado com um conjunto de resumos de referência. Nesse processo, aspectos comuns em resumos escritos por seres humanos, como sinonímia e paráfrase, não são considerados. Essa limitação é ainda mais acentuada, durante os experimentos realizados com os corpora do DUC, já que esses possuem resumos de referência abstrativos. Apesar do desenvolvimento de novos mecanismos de avaliação de sistemas de sumarização, como o *Pyramid* (NENKOVA; PASSONNEAU; MCKEOWN, 2007), eles ainda demandam uma grande carga de trabalho manual. Apesar

de ser um processo essencial para o progresso da área, a avaliação de resumos criados automaticamente ainda representa um grande desafio para a SAT.

7.4 Trabalhos Futuros

Com base nas limitações identificadas e nas lições aprendidas, como trabalhos futuros são sugeridas as seguintes linhas de investigação para a extensão e melhoria das abordagens propostas.

Novas formas de representação e ponderação de conceitos. A maioria das abordagens baseadas em conceitos usando PLI adotam unigramas (CAO et al., 2015; WAN et al., 2015) ou bigramas (GILLICK et al., 2009; BOUDIN; MOUGARD; FAVRE, 2015) como conceitos. Neste trabalho, inspirado pela investigação conduzida em (SCHLUTER; SØGAARD, 2015), outras formas de representação para a noção de conceitos também foram avaliadas: Entidades nomeadas e dependência sintáticas rotuladas e genéricas. Apesar dos bons resultados obtidos, especialmente adotando bigramas, essas formas de representação são descontextualizadas, ou seja, ações realizadas ou sofridas pelos conceitos, ou o relacionamento entre dois ou mais conceitos não são levados em consideração. Por isso, uma interessante linha de pesquisa futura é investigar a adoção de outras formas de representação mais contextualizadas, por exemplo, adotando eventos (MARUJO et al., 2016) ou estruturas na forma de triplas (sujeito, predicado e objeto) que podem ser extraídas usando sistemas de extração de relações abertas (FADER; SODERLAND; ETZIONI, 2011). Além disso, outras formas de ponderação mais adequadas para essas novas representações precisam ser investigadas. Ponderar a relevância dos conceitos extraídos de forma supervisionada aplicando algoritmos de regressão para estimar os pesos dos conceitos, também pode ser interessante para melhorar os resultados obtidos.

Técnicas de compressão e fusão de sentenças. Pesquisas envolvendo o desenvolvimento de abordagens de sumarização abstrativas têm crescido nos últimos anos. Esses sistemas adotam, principalmente, técnicas de compressão e fusão de sentenças. O sistema ICSISumm proposto por Gillick et al. (2009) aplica um processo de compressão de sentenças para eliminar partes não relevantes de uma frase analisando sua árvore sintática. Banerjee, Mitra e Sugiyama (2015b) propôs uma abordagem usando PLI que adota um método de fusão de sentenças para gerar uma única frase contendo as informações mais relevantes de duas ou mais sentenças. Trabalhos adotando técnicas de compressão e fusão de sentenças têm apresentado bom desempenho, gerando resumos menores e mais informativos do que sistemas de sumarização extrativos. A integração desses tipos de métodos tornaria as abordagens propostas em abstrativas, possibilitando assim, a geração de resumos mais informativos. Além disso, com novos resumos candidatos, possivelmente mais informativos sendo gerados, a abordagem usando PLI e regressão também se beneficiária.

Seleção dinâmica de algoritmos de regressão e características. A abordagem

baseada em conceitos usando PLI e regressão proposta foi avaliada usando apenas um único algoritmo para estimar a cobertura de informações relevantes dos resumos candidatos gerados. Trabalhos nas áreas de seleção e combinação dinâmica de algoritmos de regressão (MENDES-MOREIRA et al., 2012; REN; ZHANG; SUGANTHAN, 2016) têm demonstrado obter resultados melhores do que aplicar estaticamente somente um único algoritmo para todo objeto de teste. Neste sentido, a abordagem proposta se beneficiária de uma estratégia dinâmica de seleção que identificasse qual o algoritmo de regressão mais adequado para estimar a medida de cobertura do ROUGE-1 de um novo resumo candidato, com base em exemplos similares no conjunto de treinamento.

Além disso, com exceção do algoritmo de regressão linear que adotou um processo de seleção de características, nenhum outro algoritmo de seleção de atributos foi investigado. Estratégias de seleção independentes de algoritmo de regressão, como a filtragem baseada na correlação das características pode diminuir o desempenho da etapa de seleção do resumo mais informativo. Como ficou demonstrado com os resultados obtidos, nem sempre o algoritmo que apresentava a maior correlação e o menor erro também é o algoritmo com melhor desempenho com base nas medidas de cobertura do ROUGE. Dessa forma, selecionar as características somente com base na sua correlação pode deteriorar os resultados dos algoritmos com base nas medidas do ROUGE. Por outro lado, estratégias de seleção do tipo *Wrapper*, que são dependentes do algoritmo de regressão, em geral, possuem a desvantagem de serem muito custosos e suscetíveis a *overfitting*, ou seja, um ajuste demasiado do modelo gerado, somente para o conjunto de treinamento e ineficiente para um novo conjunto de teste. Dessa forma, uma possível linha de pesquisa futura é propor uma estratégia de seleção de atributos usando *Wrapper*, mas adotando sempre um conjunto de validação variável de modo a garantir a generalização do modelo de regressão gerado, adotando a medida do ROUGE como critério de seleção.

Adaptação das abordagens proposta para artigos de notícias escritas em Português do Brasil. Pesquisas envolvendo documentos escritos em Português do Brasil ainda são muito poucas em relação a outros idiomas, como o Inglês. Diante disso, vislumbra-se adaptar as abordagens propostas para realizar o processo de sumarização monodocumento e multidocumento, em artigos de notícias escritos em Português. Para avaliar as adaptações realizadas, experimentos podem ser realizados adotando o corpus *CST-News* (CARDOSO et al., 2011), que é muito adotado em pesquisas envolvendo artigos de notícias escritas em Português.

REFERÊNCIAS

ABUOBIEDA, A.; SALIM, N.; ALBAHAM, A. T.; OSMAN, A. H.; KUMAR, Y. J. Text summarization features selection method using pseudo genetic-based model. In: *International Conference on Information Retrieval Knowledge Management*. [S.l.: s.n.], 2012. p. 193–197.

ABUOBIEDA, A.; SALIM, N.; KUMAR, Y. J.; OSMAN, A. H. An improved evolutionary algorithm for extractive text summarization. In: SELAMAT, A.; NGUYEN, N. T.; HARON, H. (Ed.). *Intelligent Information and Database Systems*. [S.l.]: Springer Berlin Heidelberg, 2013, (Lecture Notes in Computer Science, v. 7803). p. 78–89. ISBN 978-3-642-36542-3.

AHA, D.; KIBLER, D. Instance-based learning algorithms. *Machine Learning*, v. 6, p. 37–66, 1991.

AUTOSUMMARIZER. *Automatic Text Summarizer*. 2016. Http://autosummarizer.com/. Last acess: March 2015.

BANERJEE, S.; MITRA, P.; SUGIYAMA, K. Abstractive meeting summarization using dependency graph fusion. In: *Proceedings of the 24th International Conference on World Wide Web*. New York, NY, USA: ACM, 2015. (WWW '15 Companion), p. 5–6. ISBN 978-1-4503-3473-0.

BANERJEE, S.; MITRA, P.; SUGIYAMA, K. Multi-document abstractive summarization using ilp based multi-sentence compression. In: *Proceedings of the 24th International Conference on Artificial Intelligence*. [S.l.]: AAAI Press, 2015. (IJCAI'15), p. 1208–1214. ISBN 978-1-57735-738-4.

BARALIS, E.; CAGLIERO, L.; JABEEN, S.; FIORI, A.; SHAH, S. Multi-document summarization based on the yago ontology. *Expert Syst. Appl.*, v. 40, n. 17, p. 6976–6984, 2013.

BARRERA, A.; VERMA, R. Combining syntax and semantics for automatic extractive single-document summarization. In: GELBUKH, A. F. (Ed.). *CICLing (2)*. [S.l.]: Springer, 2012. (Lecture Notes in Computer Science, v. 7182), p. 366–377. ISBN 978-3-642-28600-1.

BARZILAY, R.; ELHADAD, N.; MCKEOWN, K. R. Inferring strategies for sentence ordering in multidocument news summarization. *J. Artif. Int. Res.*, AI Access Foundation, USA, v. 17, n. 1, p. 35–55, ago. 2002. ISSN 1076-9757.

BARZILAY, R.; LAPATA, M. Modeling local coherence: An entity-based approach. *Comput. Linguist.*, MIT Press, Cambridge, MA, USA, v. 34, n. 1, p. 1–34, mar. 2008. ISSN 0891-2017.

BATISTA, J.; FERREIRA, R.; OLIVEIRA, H.; FERREIRA, R.; LINS, R. D.; SILVA, G. Pereira e; SIMSKE, S. J.; RISS, M. A quantitative and qualitative assessment of automatic text summarization systems. In: *Proceedings of the Document Engineering*. [S.l.: s.n.], 2015. (DocEng '15).

BINWAHLAN, M. S.; SALIM, N.; SUANMALI, L. Swarm based features selection for text summarization. *International Journal of Computer Science and Network Security*, p. 175–179, 2009.

BOLLEGALA, D.; OKAZAKI, N.; ISHIZUKA, M. A preference learning approach to sentence ordering for multi-document summarization. *Inf. Sci.*, Elsevier Science Inc., New York, NY, USA, v. 217, p. 78–95, dec 2012. ISSN 0020-0255.

BOUDIN, F.; MOUGARD, H.; FAVRE, B. Concept-based summarization using integer linear programming: From concept pruning to multiple optimal solutions. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, 2015. p. 1914–1918.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 30, n. 1-7, p. 107–117, abr. 1998. ISSN 0169-7552.

BROOMHEAD, D. S.; LOWE, D. Radial Basis Functions, Multi-Variable Functional Interpolation and Adaptive Networks. *Complex Systems*, v. 2, p. 321–355, mar. 1988.

CAMARGO, J. E.; GONZáLEZ, F. A. Multimodal latent topic analysis for image collection summarization. *Information Sciences*, v. 328, p. 270 – 287, 2016. ISSN 0020-0255.

CAO, Z.; WEI, F.; DONG, L.; LI, S.; ZHOU, M. Ranking with recursive neural networks and its application to multi-document summarization. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*. [S.l.: s.n.], 2015. p. 2153–2159.

CARDOSO, P.; MAZIERO, E.; JORGE, M.; SENO, E.; FELIPPO, A. D.; RINO, L.; NUNES, M.; PARDO, T. Cstnews - a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In: *Proceedings of the 3rd RST Brazilian Meeting*. [S.l.: s.n.], 2011. p. 88–105.

CESSIE, S. le; HOUWELINGEN, J. van. Ridge estimators in logistic regression. *Applied Statistics*, v. 41, n. 1, p. 191–201, 1992.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, v. 16, p. 321–357, 2002.

CHRISTENSEN, J.; MAUSAM; SODERLAND, S.; ETZIONI, O. Towards coherent multi-document summarization. In: *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*. [S.l.: s.n.], 2013. p. 1163–1173.

CONROY, J. M.; SCHLESINGER, J. D.; GOLDSTEIN, J.; O'LEARY, D. P. Left-brain/right-brain multi-document summarization. *Proceedings of the Document Understanding Conference (DUC 2004)*, 2004.

CONROY, J. M.; SCHLESINGER, J. D.; O'LEARY, D. P. Topic-focused multi-document summarization using an approximate oracle score. In: *Proceedings of the COLING/ACL on Main Conference Poster Sessions*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006. (COLING-ACL '06), p. 152–159.

DURRETT, G.; BERG-KIRKPATRICK, T.; KLEIN, D. Learning-based single-document summarization with compression and anaphoricity constraints. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. [S.l.: s.n.], 2016.

EARL, L. L. Experiments in automatic extracting and indexing. *Information Storage and Retrieval*, v. 6, n. 4, p. 313 – 330, 1970. ISSN 0020-0271.

EDMUNDSON, H. P. New methods in automatic extracting. *J. ACM*, ACM, New York, NY, USA, v. 16, n. 2, p. 264–285, abr. 1969. ISSN 0004-5411.

EDMUNDSON, H. P.; WYLLYS, R. E. Automatic abstracting and indexing - survey and recommendations. *Commun. ACM*, ACM, New York, NY, USA, v. 4, n. 5, p. 226–234, maio 1961. ISSN 0001-0782.

ENDRES-NIGGEMEYER, B. A procedural model of abstracting, and some ideas for its implementation. In: *Terminology and Knowledge Engineering (Vol. 1)*. [S.l.: s.n.], 1990. p. 230–243.

ENDRES-NIGGEMEYER, B. A naturalistic model of abstracting. *Preprints of Summarizing Text for Intelligent Communication. Dagstuhl Seminar Report*, v. 79, n. 0, 1993.

ENDRES-NIGGEMEYER, B.; HOBBS, J.; JONES, K. S. *Summarizing text for intelligent communication*. [S.l.]: Internationales Begegnungs-und Forschungszentrum für Informatik, 1993.

ERKAN, G.; RADEV, D. R. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, AI Access Foundation, USA, v. 22, n. 1, p. 457–479, dez. 2004. ISSN 1076-9757.

ETZIONI, O.; FADER, A.; CHRISTENSEN, J.; SODERLAND, S.; MAUSAM, M. Open information extraction: The second generation. In: *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume One*. [S.l.]: AAAI Press, 2011. (IJCAI'11), p. 3–10. ISBN 978-1-57735-513-7.

FADER, A.; SODERLAND, S.; ETZIONI, O. Identifying relations for open information extraction. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. (EMNLP '11), p. 1535–1545. ISBN 978-1-937284-11-4.

FATTAH, M. A hybrid machine learning model for multi-document summarization. *Applied Intelligence*, Springer US, v. 40, n. 4, p. 592–600, 2014. ISSN 0924-669X.

FATTAH, M. A.; REN, F. Ga, mr, ffnn, pnn and gmm based models for automatic text summarization. *Computer Speech and Language*, v. 23, n. 1, p. 126–144, 2009.

FERREIRA, R.; CABRAL, L. de S.; LINS, R. D.; SILVA, G. Pereira e; FREITAS, F.; CAVALCANTI, G. D.; LIMA, R.; SIMSKE, S. J.; FAVARO, L. Assessing sentence scoring techniques for extractive text summarization. *Expert Systems With Applications*, v. 40, n. 14, p. 5755–5764, 2013.

FERREIRA, R.; FREITAS, F. L. G. de; CABRAL, L. de S.; LINS, R. D.; LIMA, R.; SILVA, G. de França Pereira e; SIMSKE, S. J.; FAVARO, L. A context based text summarization system. In: *11th IAPR International Workshop on Document Analysis Systems, DAS 2014, Tours, France, April 7-10, 2014*. [S.l.: s.n.], 2014. p. 66–70.

FILIPPOVA, K. Multi-sentence compression: Finding shortest paths in word graphs. In: *COLING'10*. [S.l.: s.n.], 2010. p. 322–330.

FRANK, E. *Fully Supervised Training of Gaussian Radial Basis Function Networks in WEKA*. [S.l.]: Department of Computer Science, University of Waikato, 2014. (Working paper series (University of Waikato. Department of Computer Science)).

FREUND, Y.; SCHAPIRE, R. E. Experiments with a new boosting algorithm. In: *Thirteenth International Conference on Machine Learning*. San Francisco: Morgan Kaufmann, 1996. p. 148–156.

GAMBHIR, M.; GUPTA, V. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, p. 1–66, 2016.

GARCÍA-HERNÁNDEZ, R. A.; LEDENEVA, Y. Pattern recognition: 5th mexican conference, mcpr 2013, querétaro, mexico, june 26-29, 2013. proceedings. In: _____. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. cap. Single Extractive Text Summarization Based on a Genetic Algorithm, p. 374–383.

GIANNAKOPOULOS, G.; CONROY, J.; KUBINA, J.; RANKEL, P. A.; LLORET, E.; STEINBERGER, J.; LITVAK, M.; FAVRE, B. Multiling 2017 overview. In: *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*. [S.l.]: Association for Computational Linguistics, 2017. p. 1–6.

GIBBONS, J. D.; CHAKRABORTI, S. *Nonparametric Statistical Inference (Statistics: a Series of Textbooks and Monogrphs)*. 4. ed. [S.l.]: CRC, 2003. Hardcover. ISBN 0824740521.

GILLICK, D.; FAVRE, B.; HAKKANI-TÜR, D.; BOHNET, B.; LIU, Y.; XIE, S. The ICSI/UTD summarization system at TAC 2009. In: *Proceedings of the Second Text Analysis Conference, TAC 2009, Gaithersburg, Maryland, USA, November 16-17, 2009*. [S.l.: s.n.], 2009.

GOLDBERG, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*. 1st. ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989. ISBN 0201157675.

GROSZ, B. J.; WEINSTEIN, S.; JOSHI, A. K. Centering: A framework for modeling the local coherence of discourse. *Comput. Linguist.*, MIT Press, Cambridge, MA, USA, v. 21, n. 2, p. 203–225, jun. 1995. ISSN 0891-2017.

GUINAUDEAU, C.; STRUBE, M. Graph-based local coherence modeling. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 2013.

GUO, Z.; GAO, L.; ZHEN, X.; ZOU, F.; SHEN, F.; ZHENG, K. Spatial and temporal scoring for egocentric video summarization. *Neurocomputing*, p. –, 2016. ISSN 0925-2312.

GUPTA, P.; PENDLURI, V.; VATS, I. Summarizing text by ranking text units according to shallow linguistic features. In: *Advanced Communication Technology (ICACT), 2011 13th International Conference on*. [S.l.: s.n.], 2011. p. 1620–1625. ISSN 1738-9445.

HAGHIGHI, A.; VANDERWENDE, L. Exploring content models for multi-document summarization. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. (NAACL '09), p. 362–370. ISBN 978-1-932432-41-1.

HALL, M. A. *Correlation-based feature selection for machine learning*. [S.l.], 1998.

HAQUE, R.; NASKAR, S. K.; WAY, A.; COSTA-JUSSA, M. R.; BANCHS, R. E. Sentence similarity-based source context modelling in pbsmt. In: *Proceedings of the 2010 International Conference on Asian Language Processing*. [S.l.]: IEEE Computer Society, 2010. p. 257–260. ISBN 978-0-7695-4288-1.

HAYKIN, S. *Neural Networks: A Comprehensive Foundation*. 2nd. ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1998. ISBN 0132733501.

HIRAO, T.; YOSHIDA, Y.; NISHINO, M.; YASUDA, N.; NAGATA, M. Single-document summarization as a tree knapsack problem. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, 2013. p. 1515–1520.

HONG, K.; CONROY, J. M.; FAVRE, B.; KULESZA, A.; LIN, H.; NENKOVA, A. A repository of state of the art and competitive baseline summaries for generic news summarization. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014*. [S.l.: s.n.], 2014. p. 1608–1616.

HONG, K.; MARCUS, M.; NENKOVA, A. System combination for multi-document summarization. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2015. p. 107–117.

JOACHIMS, T. Optimizing search engines using clickthrough data. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2002. (KDD '02), p. 133–142.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in bayesian classifiers. In: *Eleventh Conference on Uncertainty in Artificial Intelligence*. San Mateo: Morgan Kaufmann, 1995. p. 338–345.

JONES, K. S. What might be in a summary? In: *Information Retrieval*. [S.l.: s.n.], 1993. p. 9–26.

JORGE, M. L. d. R. C.; PARDO, T. A. S. Experiments with cst-based multidocument summarization. In: *Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. (TextGraphs-5), p. 74–82. ISBN 978-1-932432-77-0. Disponível em: <http://dl.acm.org/citation.cfm?id=1870490.1870502>.

JUNG, W.; KO, Y.; SEO, J. Automatic text summarization using two-step sentence extraction. In: MYAENG, S.-H.; ZHOU, M.; WONG, K.-F.; ZHANG, H. (Ed.). *AIRS*. [S.l.]: Springer, 2005. (Lecture Notes in Computer Science, v. 3411), p. 71–81. ISBN 3-540-25065-4.

KHAN, A.; SALIM, N.; KUMAR, Y. J. A framework for multi-document abstractive summarization based on semantic role labelling. *Applied Soft Computing*, v. 30, p. 737 – 747, 2015. ISSN 1568-4946.

KIKUCHI, Y.; HIRAO, T.; TAKAMURA, H.; OKUMURA, M.; NAGATA, M. Single document summarization based on nested tree structure. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland: Association for Computational Linguistics, 2014. p. 315–320.

KOZIEL, S.; LEIFSSON, L.; LEES, M.; V.KRZHIZHANOVSKAYA, V.; DONGARRA, J.; SLOOT, P. M.; SEMAN, N.; JAMIL, N. Blending sentence optimization weights of unsupervised approaches for extractive speech summarization. *Procedia Computer Science*, v. 51, p. 620 – 629, 2015. ISSN 1877-0509.

KULLBACK, S. Information theory and statistic. Wiley, 1959.

KUPIEC, J.; PEDERSEN, J.; CHEN, F. A trainable document summarizer. In: *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 1995. (SIGIR '95), p. 68–73. ISBN 0-89791-714-6.

LAPATA, M. Probabilistic text structuring: Experiments with sentence ordering. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. [S.l.], 2003. p. 545–552.

LEE, G. H.; LEE, K. J. Automatic text summarization using reinforcement learning with embedding features. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. [S.l.]: Asian Federation of Natural Language Processing, 2017. p. 193–197.

LEE, H.; CHANG, A.; PEIRSMAN, Y.; CHAMBERS, N.; SURDEANU, M.; JURAFSKY, D. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Comput. Linguist.*, MIT Press, Cambridge, MA, USA, v. 39, n. 4, p. 885–916, dez. 2013. ISSN 0891-2017.

LEE, L. Measures of distributional similarity. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 1999. (ACL '99), p. 25–32. ISBN 1-55860-609-3.

LI, C.; LIU, Y.; ZHAO, L. Using external resources and joint learning for bigram weighting in ilp-based multi-document summarization. In: MIHALCEA, R.; CHAI, J. Y.; SARKAR, A. (Ed.). *HLT-NAACL*. [S.l.]: The Association for Computational Linguistics, 2015. p. 778–787. ISBN 978-1-941643-49-5.

LI, C.; QIAN, X.; LIU, Y. Using supervised bigram-based ilp for extractive summarization. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, 2013. p. 1004–1013.

LIN, C.-Y. Rouge: A package for automatic evaluation of summaries. In: MOENS, S. S. M.-F. (Ed.). *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 74–81.

LIN, J. Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theor.*, IEEE Press, Piscataway, NJ, USA, v. 37, n. 1, p. 145–151, set. 2006. ISSN 0018-9448.

LLORET, E.; PALOMAR, M. Text summarisation in progress: a literature review. *Artificial Intelligence Review*, Springer Netherlands, v. 37, n. 1, p. 1–41, 2012. ISSN 0269-2821.

LLORET, E.; PLAZA, L.; AKER, A. The challenging task of summary evaluation: an overview. *Language Resources and Evaluation*, Sep 2017. ISSN 1574-0218.

LOTHIAN, N. *Classifier4J*. 2003. Http://classifier4j.sourceforge.net/. Last acess: March 2015.

LOUIS, A.; NENKOVA, A. Automatically evaluating content selection in summarization without human models. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. (EMNLP '09), p. 306–314. ISBN 978-1-932432-59-6.

LUHN, H. P. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, IBM Corp., Riverton, NJ, USA, v. 2, n. 2, p. 159–165, abr. 1958. ISSN 0018-8646.

MANI, I. Summarization evaluation: An overview. In: *Proceedings of the Third Second Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization, NTCIR-2, Tokyo, Japan, March 7-9, 2001*. [S.l.: s.n.], 2001.

MANI, I.; KLEIN, G.; HOUSE, D.; HIRSCHMAN, L.; FIRMIN, T.; SUNDHEIM, B. Summac: A text summarization evaluation. *Nat. Lang. Eng.*, Cambridge University Press, New York, NY, USA, v. 8, n. 1, p. 43–68, mar. 2002. ISSN 1351-3249.

MANN, W. C.; THOMPSON, S. A. Rhetorical structure theory: Description and construction of text structures. In: _____. *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics*. Dordrecht: Springer Netherlands, 1987. p. 85–95. ISBN 978-94-009-3645-4.

MANN, W. C.; THOMPSON, S. A. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, v. 8, n. 3, p. 243–281, 1988.

MANNING, C. D.; SURDEANU, M.; BAUER, J.; FINKEL, J.; BETHARD, S. J.; MCCLOSKY, D. The Stanford CoreNLP natural language processing toolkit. In: *Association for Computational Linguistics (ACL) System Demonstrations*. [S.l.: s.n.], 2014. p. 55–60.

MARUJO, L.; LING, W.; RIBEIRO, R.; GERSHMAN, A.; CARBONELL, J.; MATOS, D. M. de; NETO, J. P. Exploring events and distributed representations of text in multi-document summarization. *Knowledge-Based Systems*, v. 94, p. 33 – 42, 2016. ISSN 0950-7051.

MAUSAM; SCHMITZ, M.; BART, R.; SODERLAND, S.; ETZIONI, O. Open language learning for information extraction. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. (EMNLP-CoNLL '12), p. 523–534.

MAZIERO, E. G.; JORGE, M. L. del R. C.; PARDO, T. A. S. Revisiting cross-document structure theory for multi-document discourse parsing. *Inf. Process. Manage.*, v. 50, n. 2, p. 297–314, 2014.

MCDONALD, R. A study of global inference algorithms in multi-document summarization. In: *Proceedings of the 29th European Conference on IR Research*. Berlin, Heidelberg: Springer-Verlag, 2007. (ECIR'07), p. 557–564. ISBN 978-3-540-71494-1.

MEENA, Y.; DEOLIA, P.; GOPALANI, D. Optimal features set for extractive automatic text summarization. In: *Advanced Computing Communication Technologies (ACCT), 2015 Fifth International Conference on*. [S.l.: s.n.], 2015. p. 35–40.

MEENA, Y. K.; GOPALANI, D. Analysis of sentence scoring methods for extractive automatic text summarization. In: *Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies*. New York, NY, USA: ACM, 2014. (ICTCS '14), p. 53:1–53:6. ISBN 978-1-4503-3216-3.

MENDES-MOREIRA, J. a.; SOARES, C.; JORGE, A. M.; SOUSA, J. F. D. Ensemble approaches for regression: A survey. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 45, n. 1, p. 10:1–10:40, dez. 2012. ISSN 0360-0300.

MENDOZA, M.; BONILLA, S.; NOGUERA, C.; COBOS, C.; LEóN, E. Extractive single-document summarization based on genetic operators and guided local search. *Expert System with Applications*, v. 41, n. 9, p. 4158–4169, 2014.

MIHALCEA, R.; TARAU, P. Textrank: Bringing order into texts. In: LIN, D.; WU, D. (Ed.). *Proceedings of EMNLP 2004*. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 404–411.

MILLER, G. A. Wordnet: A lexical database for english. *Commun. ACM*, ACM, New York, NY, USA, v. 38, n. 11, p. 39–41, nov. 1995. ISSN 0001-0782.

MNIH, V.; KAVUKCUOGLU, K.; SILVER, D.; RUSU, A. A.; VENESS, J.; BELLEMARE, M. G.; GRAVES, A.; RIEDMILLER, M.; FIDJELAND, A. K.; OSTROVSKI, G.; PETERSEN, S.; BEATTIE, C.; SADIK, A.; ANTONOGLOU, I.; KING, H.; KUMARAN, D.; WIERSTRA, D.; LEGG, S.; HASSABIS, D. Human-level control through deep reinforcement learning. *Nature*, Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., v. 518, n. 7540, p. 529–533, feb 2015. ISSN 00280836.

NENKOVA, A. Automatic text summarization of newswire: Lessons learned from the document understanding conference. In: *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA*. [S.l.: s.n.], 2005. p. 1436–1441.

NENKOVA, A.; MCKEOWN, K. A survey of text summarization techniques. In: AGGARWAL, C. C.; ZHAI, C. (Ed.). *Mining Text Data*. [S.l.]: Springer, 2012. p. 43–76. ISBN 978-1-4419-8462-3.

NENKOVA, A.; PASSONNEAU, R.; MCKEOWN, K. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, ACM, New York, NY, USA, v. 4, n. 2, maio 2007. ISSN 1550-4875.

NENKOVA, A.; VANDERWENDE, L.; MCKEOWN, K. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In: EFTHIMIADIS, E. N.; DUMAIS, S. T.; HAWKING, D.; JäRVELIN, K. (Ed.). *SIGIR*. [S.l.]: ACM, 2006. p. 573–580. ISBN 1-59593-369-7.

NETO, J. L.; FREITAS, A. A.; KAESTNER, C. A. A. Automatic text summarization using a machine learning approach. In: BITTENCOURT, G.; RAMALHO, G. (Ed.). *Advances in Artificial Intelligence*. [S.l.]: Springer Berlin Heidelberg, 2002, (Lecture Notes in Computer Science, v. 2507). p. 205–215. ISBN 978-3-540-00124-9.

OLIVEIRA, H.; FERREIRA, R.; LIMA, R.; LINS, R. D.; FREITAS, F.; RISS, M.; SIMSKE, S. J. Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization. *Expert Syst. Appl.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 65, n. C, p. 68–86, dez. 2016a. ISSN 0957-4174.

OUYANG, Y.; LI, W.; LU, Q.; ZHANG, R. A study on position information in document summarization. In: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. (COLING '10), p. 919–927.

OVER, P.; DANG, H.; HARMAN, D. Duc in context. *Information Process and Management*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 43, n. 6, p. 1506–1520, nov. 2007. ISSN 0306-4573.

OWCZARZAK, K.; CONROY, J. M.; DANG, H. T.; NENKOVA, A. An assessment of the accuracy of automatic evaluation in summarization. In: *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. p. 1–9.

OWCZARZAK, K.; DANG, H. T. Overview of the TAC 2011 summarization track: Guided task and AESOP task. In: *Proceedings of the Text Analysis Conference (TAC 2011), Gaithersburg, Maryland, USA, November*. [S.l.: s.n.], 2011.

PAICE, C. D. The automatic generation of literature abstracts: An approach based on the identification of self-indicating phrases. In: *Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval*. Kent, UK, UK: [s.n.], 1981. (SIGIR '80), p. 172–191. ISBN 0-408-10775-8.

PALSHIKAR, G. K.; DESHPANDE, S.; ATHIAPPAN, G. Combining summaries using unsupervised rank aggregation. In: GELBUKH, A. F. (Ed.). *CICLing (2)*. [S.l.]: Springer, 2012. (Lecture Notes in Computer Science, v. 7182), p. 378–389. ISBN 978-3-642-28600-1.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, Now Publishers Inc., Hanover, MA, USA, v. 2, n. 1-2, p. 1–135, jan. 2008. ISSN 1554-0669.

PARVEEN, D.; RAMSL, H.; STRUBE, M. Topical coherence for graph-based extractive summarization. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. [S.l.: s.n.], 2015a. p. 1949–1954.

PARVEEN, D.; STRUBE, M. Integrating importance, non-redundancy and coherence in graph-based extractive summarization. In: *Proceedings of the 24th International Conference on Artificial Intelligence*. [S.l.]: AAAI Press, 2015b. (IJCAI'15), p. 1298–1304. ISBN 978-1-57735-738-4.

PEI, Y.; YIN, W.; FAN, Q.; HUANG, L. A supervised aggregation framework for multi-document summarization. In: *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*. [S.l.: s.n.], 2012. p. 2225–2242.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. [S.l.: s.n.], 2014. p. 1532–1543.

PEYRARD, M.; ECKLE-KOHLER, J. A general optimization framework for multi-document summarization using genetic algorithms and swarm intelligence. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, 2016. p. 247–257.

PITLER, E.; RAGHUPATHY, M.; MEHTA, H.; NENKOVA, A.; LEE, A.; JOSHI, A. K. Easily identifiable discourse relations. In: *COLING 2008, 22nd International Conference on Computational Linguistics, Posters Proceedings, 18-22 August 2008, Manchester, UK*. [S.l.: s.n.], 2008. p. 87–90.

PLATT, J. Fast training of support vector machines using sequential minimal optimization. In: SCHOELKOPF, B.; BURGES, C.; SMOLA, A. (Ed.). *Advances in Kernel Methods - Support Vector Learning*. [S.l.]: MIT Press, 1998.

PORTER, M. F. Readings in information retrieval. In: JONES, K. S.; WILLETT, P. (Ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997. cap. An Algorithm for Suffix Stripping, p. 313–316. ISBN 1-55860-454-5.

QUINLAN, J. R. Learning with continuous classes. In: . [S.l.]: World Scientific, 1992. p. 343–348.

QUINLAN, R. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

RANKEL, P. A.; CONROY, J. M.; SLUD, E.; O'LEARY, D. P. Ranking human and machine summarization systems. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*. [S.l.: s.n.], 2011. p. 467–473.

RATH, G. J.; RESNICK, A.; SAVAGE, T. R. The formation of abstracts by the selection of sentences. part i. sentence selection by men and machines. *American Documentation*, Wiley Subscription Services, Inc., A Wiley Company, v. 12, n. 2, p. 139–141, 1961. ISSN 1936-6108.

RAU, L. F.; JACOBS, P. S.; ZERNIK, U. Information extraction and text summarization using linguistic knowledge acquisition. *Information Processing & Management*, v. 25, n. 4, p. 419 – 428, 1989. ISSN 0306-4573.

REN, P.; CHEN, Z.; REN, Z.; WEI, F.; MA, J.; RIJKE, M. de. Leveraging contextual sentence relations for extractive summarization using a neural attention model. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2017. (SIGIR '17), p. 95–104. ISBN 978-1-4503-5022-8.

REN, P.; WEI, F.; CHEN, Z.; MA, J.; ZHOU, M. A redundancy-aware sentence regression framework for extractive summarization. In: *COLING*. [S.l.]: ACL, 2016. p. 33–43.

REN, Y.; ZHANG, L.; SUGANTHAN, P. N. Ensemble classification and regression-recent developments, applications and future directions [review article]. *IEEE Computational Intelligence Magazine*, v. 11, n. 1, p. 41–53, Feb 2016. ISSN 1556-603X.

RESNIK, P. Using information content to evaluate semantic similarity in a taxonomy. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. (IJCAI'95), p. 448–453. ISBN 1-55860-363-8, 978-1-558-60363-9.

ROUSSEEUW, P. J. Least median of squares regression. *Journal of the American Statistical Association*, Taylor and Francis, v. 79, n. 388, p. 871–880, 1984.

RUSH, J. E.; SALVADOR, R.; ZAMORA, A. Automatic abstracting and indexing. ii. production of indicative abstracts by application of contextual inference and syntactic coherence criteria. *Journal of the American Society for Information Science*, Wiley Subscription Services, Inc., A Wiley Company, v. 22, n. 4, p. 260–274, 1971. ISSN 1097-4571.

SAGGION, H.; POIBEAU, T. Automatic text summarization: Past, present and future. In: POIBEAU, T.; SAGGION, H.; PISKORSKI, J.; YANGARBER, R. (Ed.). *Multi-source, Multilingual Information Extraction and Summarization*. [S.l.]: Springer Berlin Heidelberg, 2013, (Theory and Applications of Natural Language Processing). p. 3–21. ISBN 978-3-642-28568-4.

SAGGION, H.; TORRES-MORENO, J.-M.; CUNHA, I. d.; SANJUAN, E. Multilingual summarization evaluation without human models. In: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. (COLING '10), p. 1059–1067.

SALTON, G.; SINGHAL, A.; MITRA, M.; BUCKLEY, C. Automatic text structuring and summarization. *Information Processing & Management*, v. 33, n. 2, p. 193 – 207, 1997. ISSN 0306-4573.

SCHLUTER, N.; SØGAARD, A. Unsupervised extractive summarization via coverage maximization with syntactic and semantic concepts. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China: Association for Computational Linguistics, 2015. p. 840–844.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, v. 52, n. 3/4, p. 591–611, Dec. 1965.

SHEVADE, S.; KEERTHI, S.; BHATTACHARYYA, C.; MURTHY, K. Improvements to the smo algorithm for svm regression. In: *IEEE Transactions on Neural Networks*. [S.l.: s.n.], 1999.

SHIRKHORSHIDI, A. S.; AGHABOZORGI, S.; WAH, T. Y. A comparison study on similarity and dissimilarity measures in clustering continuous data. *PLOS ONE*, Public Library of Science, v. 10, n. 12, p. 1–20, 12 2015.

SILVA, G.; FERREIRA, R.; LINS, R. D.; CABRAL, L.; OLIVEIRA, H.; SIMSKE, S. J.; RISS, M. Automatic text document summarization based on machine learning. In: *Proceedings of the 2015 ACM Symposium on Document Engineering*. New York, NY, USA: ACM, 2015. (DocEng '15), p. 191–194. ISBN 978-1-4503-3307-8.

SILVA, G. Pereira e; FERREIRA, R.; OLIVEIRA, H.; CABRAL, L.; LINS, R. D.; SIMSKE, S. J.; RISS, M. Automatic text document summarization based on machine learning. In: *Proceedings of the Document Engineering*. [S.l.: s.n.], 2015. (DocEng '15).

SIPOS, R.; SHIVASWAMY, P.; JOACHIMS, T. Large-margin learning of submodular summarization models. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. (EACL '12), p. 224–233. ISBN 978-1-937284-19-0.

SUANMALI, L.; SALIM, N.; BINWAHLAN, M. S. Fuzzy logic based method for improving text summarization. *CoRR*, abs/0906.4690, 2009.

SUCHANEK, F. M.; KASNECI, G.; WEIKUM, G. Yago: A core of semantic knowledge. In: *Proceedings of the 16th International Conference on World Wide Web*. New York, NY, USA: ACM, 2007. (WWW '07), p. 697–706. ISBN 978-1-59593-654-7.

TAN, P.-N.; KUMAR, V.; SRIVASTAVA, J. Selecting the right interestingness measure for association patterns. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2002. (KDD '02), p. 32–41. ISBN 1-58113-567-X.

TOFFLER, A. *Future shock*. [S.l.]: Random House, 1970. ISBN 0394425863.

TORRES-MORENO, J.-M. Automatic text summarization. In: _____. *Automatic Text Summarization*. [S.l.]: John Wiley & Sons, Inc., 2014. p. 53–108. ISBN 9781119004752.

WAN, X.; CAO, Z.; WEI, F.; LI, S.; ZHOU, M. Multi-document summarization via discriminative summary reranking. *CoRR*, abs/1507.02062, 2015. Disponível em: <http://arxiv.org/abs/1507.02062>.

WAN, X.; YANG, J. Multi-document summarization using cluster-based link analysis. In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2008. (SIGIR '08), p. 299–306. ISBN 978-1-60558-164-4.

WANG, D.; LI, T. Weighted consensus multi-document summarization. *Inf. Process. Manage.*, v. 48, n. 3, p. 513–523, 2012.

WIEGAND, W. A.; JR., D. G. D. *Encyclopedia of Library History*. [S.l.]: Routledge, 1994. ISBN 0824057872.

ZAJIC, D.; DORR, B. J.; LIN, J. J.; SCHWARTZ, R. M. Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. *Inf. Process. Manage.*, v. 43, n. 6, p. 1549–1570, 2007.

Apêndices

APÊNDICE A – AVALIANDO MÉTODOS PARA A PONDERAÇÃO E FORMAS DE REPRESENTAÇÃO DE CONCEITOS PARA A SUMARIZAÇÃO MONODOCUMENTO

Recentemente, algumas das principais abordagens do estado da arte para SAT extrativa são baseadas na noção da maximização da cobertura de conceitos importantes utilizando PLI (GILLICK et al., 2009; BOUDIN; MOUGARD; FAVRE, 2015; CAO et al., 2015; LI; LIU; ZHAO, 2015). Essas abordagens baseiam-se na premissa de que resumos informativos podem ser gerados, selecionando o subconjunto de sentenças que maximiza a cobertura de conceitos relevantes do(s) documento(s) de entrada, respeitando o limiar do tamanho máximo do resumo a ser gerado. Duas questões fundamentais que precisam ser previamente definidas nesse tipo de abordagem são: **(i)** Qual a forma de representação adotar para modelar a noção de conceitos; e **(ii)** Como mensurar a relevância desses conceitos.

Diversos trabalhos (GILLICK et al., 2009; BOUDIN; MOUGARD; FAVRE, 2015; CAO et al., 2015; LI; LIU; ZHAO, 2015) presentes na literatura adotam unigramas ou bigramas como formas de representação de conceitos. Tradicionalmente, os pesos atribuídos a esses conceitos têm sido estimados: **(i)** Adotando o total de documentos que mencionam o conceito (frequência dos documentos) (GILLICK et al., 2009; BOUDIN; MOUGARD; FAVRE, 2015) para o cenário multidocumento; ou **(ii)** Combinando diversos métodos, como posição e frequência, em conjunto com algoritmos de regressão para estimar a importância dos conceitos, com base em exemplos de treinamento (CAO et al., 2015; LI; LIU; ZHAO, 2015).

No melhor do conhecimento do autor deste trabalho, nenhum esforço foi identificado na literatura, para comparar o desempenho de diferentes métodos de pontuação, e formas de representação de conceitos no contexto da sumarização monodocumento. O trabalho mais próximo encontrado foi desenvolvido por Schluter e Søgaard (2015), cujo objetivo era investigar o desempenho de cinco estratégias para a representação individuais e combinadas de conceitos. Os autores realizaram experimentos adotando três corpora nos domínios de notícias, documentos jurídicos, e artigos da Wikipédia. Para ponderar os pesos dos conceitos, os autores aplicaram o método de frequência dos conceitos nos documentos.

Neste apêndice são apresentados os experimentos realizados para avaliar cinco estratégias para representação dos conceitos, além de diferentes métodos estatísticos, e baseados em grafos para ponderar a importância desses conceitos. Os experimentos realizados, na

tarefa de sumarização monodocumento, concentram-se no domínio de artigos de notícias escritos em Inglês, permitindo assim, uma melhor generalização dos resultados do que utilizando somente um corpus por domínio. Para isso, os corpora do DUC 2001-2002, e CNN foram adotados.

O restante deste apêndice está organizado da seguinte forma: Na Seção A.1 é apresentada a abordagem baseada em conceitos adotada nos experimentos, bem como os métodos de ponderação e as formas de representação investigadas. Os resultados experimentais são apresentados e discutidos na Seção A.2. Finalmente, na Seção A.3 as conclusões dos experimentos realizados são delineadas.

A.1 Abordagem Baseada em Conceitos Utilizando Programação Linear Inteira

A abordagem baseada em conceitos proposta por Gillick et al. (2009) é centrada na ideia de selecionar o subconjunto de sentenças que maximiza a cobertura de conceitos importantes do documento de entrada, e ao mesmo tempo atenda a restrição do tamanho máximo do resumo imposta. Esse problema de otimização combinatória é modelado utilizando como função objetivo a equação $MAX \quad \sum w_i c_i$, na qual c_i representa um conceito, e w_i sua respectiva pontuação (peso) de importância. Para a resolução desse problema de máxima cobertura, Programação Linear Inteira (PLI) é adotada para encontrar soluções exatas (MCDONALD, 2007).

Gillick et al. (2009) sugere que n-gramas, entidades nomeadas, elementos sintáticos, relações semânticas, entre outras representações, podem ser adotadas como conceitos. Contudo, ele alerta que quanto mais erros forem gerados durante a extração, ou a ponderação da importância dos conceitos, pior são resultados obtidos na seleção das sentenças. Diante disso, é de suma importância definir de maneira adequada qual a forma de representação utilizar, e como ponderar a importância dos conceitos extraídos.

A abordagem baseada em conceitos utilizada nos experimentos relatados neste Apêndice é uma versão simplificada da abordagem proposta no Capítulo 4, sendo ela composta por cinco etapas brevemente descritas a seguir.

1. **Pré-processamento:** Esta primeira etapa realiza o pré-processamento dos documentos de entrada utilizando a ferramenta Stanford CoreNP (MANNING et al., 2014). As tarefas de PLN executadas são: tokenização, segmentação das sentenças, atribuição das classes gramaticais das palavras, lematização, análise de dependência, e extração de entidades nomeadas.

2. **Extração dos Conceitos:** Nesta etapa, os documentos de entrada são analisados para a extração dos conceitos. Além disso, conceitos formados somente por *stopwords*

e contendo símbolos de pontuação são removidos (GILLICK et al., 2009; BOUDIN; MOUGARD; FAVRE, 2015).

3. **Ponderação dos Conceitos:** Nesta etapa, os métodos de ponderação investigados são adotados para atribuir um escore de importância para cada conceito extraído na etapa anterior.

4. **Filtragem das Sentenças:** Incluir sentenças muito pequenas nos resumos não é uma estratégia adequada, uma vez que essas sentenças, geralmente, não são representativas o suficiente para justificar a sua inclusão nos resumos. De maneira similar, selecionar sentenças muito grandes pode ser um desperdício de espaço. Dessa forma, sentenças com dez ou menos palavras ou com setenta ou mais palavras são removidas (FERREIRA et al., 2013; BOUDIN; MOUGARD; FAVRE, 2015). Além disso, sentenças duplicadas também são descartadas.

5. **Geração dos Resumos:** Esta última etapa formula a tarefa de seleção de sentenças como um problema de otimização, conforme apresentado na Equação A.1. Para solucionar tal problema, PLI foi utilizada por meio da ferramenta GLPK[1].

$$MAX \quad \sum_{c_i \in C} w_i c_i \tag{A.1a}$$

$$s.t. \quad \sum_{s_j \in S} l_j s_j \leq L \tag{A.1b}$$

$$s_j Occ_{ij} \leq c_i \quad \forall i, j \tag{A.1c}$$

$$\sum_{s_j \in S} s_j Occ_{ij} \geq c_i \quad \forall i, j \tag{A.1d}$$

$$c_i, s_j, Occ_{ij} \in \{0, 1\} \quad \forall i, j \tag{A.1e}$$

na qual,

- c_i, s_j e Occ_{ij} são variáveis binárias que indicam o conceito (c_i), a sentença (s_j), e a presença do conceito c_i na sentença s_j, respectivamente.
- w_i é a pontuação (peso) de relevância atribuída a cada conceito c_i pertencente ao conjunto de conceitos C extraído do documento de entrada.
- l é o tamanho de cada sentença s_j do conjunto de sentenças S contidas no documento;
- L é o tamanho máximo do resumo que deve ser gerado.

[1] https://www.gnu.org/software/glpk/

A Equação A.1c e a Equação A.1d garantem a consistência do modelo, ou seja, se uma frase é selecionada, isto implica em selecionar todos os conceitos que ela contém, e um conceito só é selecionado se, e somente se, ele está presente em pelo menos uma sentença selecionada.

As cinco formas de representações, e os métodos de ponderação dos conceitos avaliados neste experimento são apresentados na Subseção A.1.1 e Subseção A.1.2, respectivamente.

A.1.1 Formas de Representação dos Conceitos

As seguintes formas para representação dos conceitos sugeridas por Gillick et al. (2009) e Schluter e Søgaard (2015) são investigadas:

Unigramas e Bigramas: Essas duas formas de representações são as mais simples, e adotadas na literatura (GILLICK et al., 2009; BOUDIN; MOUGARD; FAVRE, 2015; CAO et al., 2015; LI; LIU; ZHAO, 2015).

Entidades Nomeadas: A ideia dessa representação é utilizar como conceitos expressões que se referem a nomes de pessoas, lugares, organizações, entre outros. Tais elementos são importantes para sumarização pois descrevem entidades do mundo real que são mencionados no documento.

Dependências Sintáticas: Essa representação utiliza as dependências sintáticas entre as palavras como conceitos. Algumas dessas dependências podem ser do tipo sujeito, objeto direto e indireto, complemento, entre outras. Por exemplo, dada a frase "*John walks on the beach.*", as seguintes dependências são extraídas: *root(ROOT, walks)*, *nsubj(walks, John)*, *case(beach, on)*, *det(beach,the)*, *nmod:on(walks, beach)*. Assim como proposto por Schluter e Søgaard (2015), duas formas de representação são derivadas a partir das dependências sintáticas: **(i)** Usando explicitamente o tipo da dependência para rotular a relação, por exemplo, *nsubj(walks, John)*; e **(ii)** Adotando um tipo genérico para rotular as dependências, por exemplo, *dep(walks, John)*.

A.1.2 Métodos de Ponderação dos Conceitos

Analisando os resultados dos experimentos apresentados no Capítulo 3, identificou-se que as técnicas de pontuação de sentenças superficiais que apresentam os melhores resultados são baseadas na frequência, posição, e centralidade, das informações. Diante disso, diferentes métodos que exploram esses três aspectos foram investigados para ponderar os pesos dos conceitos. Os métodos utilizados nos experimentos foram divididos em estatísticos e baseados em grafos, conforme descritos nas subseções a seguir.

A.1.3 Métodos Estatísticos

Os seguintes métodos estatísticos são avaliados neste trabalho:

Frequência do Conceito - Frequência Inversa do Conceito (FC-FIC): Este método é baseado no tradicional TF-IDF, só que aplicado em nível de sentença. A pontuação de um conceito c_i utilizando esse método é computado conforme apresentado na Equação A.2.

$$FC - FIC(c_i) = Freq(c_i) \times log\left(\frac{S}{s_{c_i}}\right) \tag{A.2}$$

- $Freq$ retorna a frequência de um conceito c_i no documento.
- S é o total de sentenças do documento.
- s_{c_i} é o total de sentenças em que c_i está presente.

Frequência das Sentenças: Neste método, a importância de um conceito é dada pela sua centralidade, em nível de sentenças, ou seja, o total de sentenças em que o conceito está presente.

Frequência do Conceito: Este método utiliza como peso de um conceito, o total de vezes que ele aparece no documento de entrada.

Posição das Sentenças: Utilizar as sentenças que aparecem no início dos documentos para compor os resumos é uma das estratégias mais antigas e eficientes para SAT (FERREIRA et al., 2013; OLIVEIRA et al., 2016a), principalmente em artigos de notícias. Para capturar essa suposição, nesse método, conceitos que aparecem mais próximos do início do documento recebem maior importância. Na Equação A.3 é apresentado como o peso de um conceito é computado utilizando este método.

$$SentPos(c_i) = 1 - \frac{Index_{s_{c_i}}}{S} \tag{A.3}$$

- $Index_{s_{c_i}}$ retorna o índice da primeira frase que menciona o conceito c_i, com a contagem sendo iniciada por 0.
- S é o total de sentenças do documento.

A.1.4 Métodos Baseados em Grafos

Para computar a pontuação dos conceitos utilizando estes métodos, primeiramente é preciso construir um grafo para representar os conceitos extraídos. Dessa forma, um grafo de conceitos é criado para cada documento de entrada, no qual os vértices são conceitos e as arestas consistem em relações de adjacência entre dois conceitos. Após a construção

do grafo, várias medidas podem ser computadas para atribuir uma pontuação para cada vértice (conceito) criado. Os métodos baseados em grafos investigados neste apêndice são descritos resumidamente a seguir:

Betweenness Centrality computa o número de caminhos mais curtos entre dois vértices do grafo de conceitos que inclui o vértice v_i. Na Equação A.4 é apresentada como essa medida é computada.

$$Betweenness(v_i) = \frac{CMCv_i}{|CMC|} \tag{A.4}$$

- $CMCv_i$ é o total de caminhos mais curtos ligando dois vértices que contém o vértice v_i.
- $|CMC|$ é o total de caminhos mais curtos computados entre todos os vértices do grafo.

Closeness centrality é a soma das distâncias mais curtas entre um vértice e todos os outros vértices do grafo. A Equação A.5 demonstra como essa medida é computada.

$$Closeness(v_i) = \frac{|V| - 1}{\sum_{i \neq j}^{V} CaminhoMaisCurto(v_i, v_j)} \tag{A.5}$$

- $CaminhoMaisCurto(v_i, v_j)$ retorna a distância do caminho mais curto entre os vértices v_i e v_j.
- V é o conjunto de vértices no grafo.
- $|V|$ é o total de vértices no grafo.

Eigenvector centrality computa a centralidade de um vértice em função das centralidades de seus vizinhos. Essa medida baseia-se no pressuposto de que vértices conectados com outros muito importantes devem ter maior peso do que aqueles ligados com vértices de menor relevância. É apresentada na Equação A.6 como essa medida é calculada.

$$Eigenvector(v_i) = \sum_{v_j \in Inc(v_i)} w_{ji} \times Eigenvector(v_j) \tag{A.6}$$

- $Inc(v_i)$ é o conjunto de vértices incidentes com o vértice v_i.
- w_{ji} é o número de coocorrências entre os vértices v_i e v_j.

Grau é o número de arestas incidentes em um vértice. No grafo de conceitos, o grau de um conceito c_i indica o número de conceitos que co-ocorrem diretamente com c_i.

Hypertext Induced Topic Search (HITS) é um popular algoritmo para análise de ligação entre páginas *Web*. O algoritmo HITS classifica cada página, que é representada como um vértice no grafo, como *Hub* ou Autoridade. Autoridade é um vértice com muitas arestas partindo dele, enquanto que um *Hub* é um vértice que recebe muitas arestas. Este trabalho utilizou uma versão ponderada desse algoritmo, em que o peso das arestas foi definido como o número de coocorrências entre os dois vértices. Na Equação A.7 é apresentada como a pontuação de um vértice é computada usando este algoritmo.

$$Autoridade(v_i) = |Out(v_i)| \tag{A.7a}$$

$$Hub(v_i) = |In(v_i)| \tag{A.7b}$$

- $|In(v_i)|$ é o total de vértices que apontam para v_i,
- $|Out(v_i)|$ é o total de vértices que v_i aponta.

PageRank é um algoritmo clássico proposto por Brin e Page (1998) para ranquear páginas *Web* com base na estrutura de suas ligações. Esse algoritmo parte da premissa de que uma página, que é um vértice no grafo, é relevante caso outras páginas importantes possuam ligações (menções) a ela. Na Equação A.8 é definida como a pontuação de um vértice é computada usando esse algoritmo.

$$PageRank(v_i) = (1 - d) + d \times \sum_{v_j \in In(v_i)} \frac{PageRank(v_j)}{|Out(v_j)|} \tag{A.8}$$

- d é um fator de amortecimento, que geralmente é definido como 0,85 (BRIN; PAGE, 1998).
- $In(v_i)$ é o conjunto de vértices que apontam para v_i,
- $|Out(v_i)|$ é o total de vértices que v_i aponta.

TextRank é um algoritmo de ranqueamento (MIHALCEA; TARAU, 2004) baseado no PageRank usado para extrair palavras chave, e determinar seus respectivos pesos usando um modelo baseado em grafos. O *TextRank* também é baseado na ideia de utilizar o peso dos vértices vizinhos para mensurar a importância de um vértice. Na Equação A.9 é apresentada como a pontuação de um vértice é computada usando este algoritmo.

$$T(v_i) = (1 - d) + d \times \sum_{v_j \in In(v_i)} \frac{w_{ji}}{\sum_{v_k \in Out(v_j)} w_{jk}} \times T(v_j) \tag{A.9}$$

- d é um fator de amortecimento, que geralmente é definido como 0.85 (BRIN; PAGE, 1998).

- *In(v_i)* é o conjunto de vértices que apontam para v_i,
- *Out(v_j)* é o conjunto de vértices que v_j aponta.
- w_{ji} é o número de coocorrências entre os vértices v_i e v_j.

A.2 Experimentos

Nesta seção são apresentados os experimentos realizados para avaliar as formas de representação e os métodos de ponderação de conceitos descritos na Seção A.1. Primeiramente, na Subseção A.2.1, os corpora adotados, as medidas de avaliação utilizadas e os limiares de sumarização usados são delineados. Os resultados dos experimentos realizados são apresentados e discutidos na Subseção A.2.2.

A.2.1 Configurações dos Experimentos

Os experimentos foram realizados usando os tradicionais corpora do DUC 2001-2002, e o corpus CNN. Detalhes desses corpora podem ser encontrados no Capítulo 4. Para avaliar os resumos gerados, as seguintes medidas quantitativas são usadas:

ROUGE-N: As medidas de cobertura do ROUGE-1 (R-1) e ROUGE-2 (R-2) (LIN, 2004) foram computadas para todos os experimentos conduzidos. A versão 1.5.5 do ROUGE foi usada com os seguintes parâmetros: *-n 2 -m -c 95 -f A*. Nos corpora do DUC, o parâmetro *-l N* foi usado para truncar todos os resumos para no máximo n palavras.

Intersecção de Sentenças (IS): Essa medida computa a intersecção de sentenças entre os resumos gerados e os resumos de referência disponíveis. Essa medida só pode ser calculada quando existem resumos de referência extrativos. Assim, a medida IS é calculada apenas para o corpus CNN.

Para o corpus CNN, a taxa de compressão de 10% do número de sentenças do documento de entrada foi adotada como limiar de sumarização. Nos corpora do DUC 2001-2002, os resumos gerados possuem no máximo 105 palavras.

A.2.2 Resultados Experimentais

Na Tabela 40 e na Tabela 41 são apresentados os resultados dos experimentos realizados no corpus CNN usando os métodos estatísticos e baseados em grafos, respectivamente. Já a Tabela 42 e a Tabela 43 resumem os resultados obtidos com as avaliações dos métodos estatísticos e baseados em grafos nos corpora do DUC 2001-2002, respectivamente. Para verificar se existem diferenças estatísticas nas comparações realizadas, o teste de *Wilcoxon signed-rank* foi utilizado, adotando um nível de 95% de confiança. O melhor resultado

global para cada corpus é destacado em negrito e o grupo de resultados estatisticamente similares, se existir, são indicados pelo símbolo (†).

Tabela 40 – Resultados (%) e desvio padrão entre parênteses das avaliações dos métodos de ponderação estatísticos no corpus CNN. O melhor desempenho global está destacado em negrito, e o grupo de configurações estatisticamente semelhantes, se existir, é indicado pelo símbolo †.

Métodos Estatísticos	Representações	CNN		
		IS	R-1	R-2
Freq. das Sentenças	Unigrama	20,67	53,54 (18,43)	31,39 (24,22)
	Bigrama	21,03	54,10 (19,32)	32,84 (25,25)
	Entidade Nomeada	18,84	46,40 (18,64)	26,09 (22,74)
	Dep. Sintática G.	19,95	53,77 (19,27)	31,75 (25,31)
	Dep. Sintática R.	19,71	53,54 (19,28)	31,43 (25,32)
Freq. do Conceito	Unigrama	20,41	53,28 (18,33)	31,05 (23,94)
	Bigrama	20,93	54,00 (19,31)	32,73 (25,16)
	Entidade Nomeada	18,88	46,42 (18,53)	26,12 (22,62)
	Dep. Sintática G.	20,06	53,84 (19,28)	31,83 (25,37)
	Dep. Sintática R.	19.86	53,67 (19,27)	31,59 (25,35)
FC-FIC	Unigrama	19,89	53,01 (18,48)	30,64 (24,29)
	Bigrama	20,81	53,92 (19,54)	32,42 (25,58)
	Entidade Nomeada	19,16	46,75 (18,81)	26,61 (22,93)
	Dep. Sintática G.	19,50	53,41 (19,43)	31,18 (25,48)
	Dep. Sintática R.	19,59	53,44 (19,47)	31,25 (25,59)
Pos. do Conceito	Unigrama	20,34	52,80 (18,38)	30,75 (24,03)
	Bigrama	20,88	53,85 (19,37)	32,48 (25,32)
	Entidade Nomeada	18,57	45,82 (18,38)	25,59 (22,43)
	Dep. Sintática G.	19,82	53,43 (19,11)	31,38 (25,11)
	Dep. Sintática R.	19,58	53,25 (19,13)	31,14 (25,09)
Pos. da Sentença	Unigrama	20,34	52,88 (18,80)	31,10 (24,83)
	Bigrama	**25,33**	**56,45** (19,52)	**37,21** (25,62)
	Entidade Nomeada	22,91	49,32 (19,54)	30,63 (24,09)
	Dep. Sintática G.	24,52	56,14† (19,64)	36,34† (25,88)
	Dep. Sintática R.	24,44	56,07† (19,50)	36,25† (25,67)

O método de posição das sentenças apresentou melhorias estatisticamente significantes em relação a todos os demais métodos de ponderação de conceitos em termos das medidas de cobertura do ROUGE-1 e ROUGE-2, nos corpora do DUC 2002 e CNN. Além disso, o método de posição de sentenças obteve os melhores resultados na medida IS no corpus CNN. No DUC 2001, o método de posição das sentenças também alcançou o melhor desempenho em todas as medidas de avaliação, mas essa diferença não foi estatisticamente superior em relação ao método de frequência do conceito adotando bigramas e dependências sintáticas sem a identificação das relações (Dep. Sintática G.), e ao método de frequência das sentenças com bigramas. Outros dois métodos que também apresentaram bom desempenho nos três corpora foram os métodos de frequência do conceito e frequência das sentenças.

Em relação às formas de representação dos conceitos, bigramas apresentou melhor desempenho em relação ás demais representações nos corpora do DUC 2002 e CNN. Somente no DUC 2001, que a estratégia de adotar as dependências sintáticas com a identificação das relações (Dep. Sintática R.) gerou melhor desempenho. Em todos os corpora, o método de posição de sentenças usando bigramas, Dep. Sintática R. ou Dep. Sintática G., produziu resultados estatisticamente similares entre eles, mas superiores em comparação com as demais formas de representação e métodos de pontuação. Analisando as 91 comparações realizadas, adotar bigramas levou a resultados superiores em relação às demais formas de representação em 63,73% dos casos, unigramas obteve a melhor performance em 18,68%, Dep. Sintática G. em 8,79%, Dep. Sintática S em 5,50%, e Entidades Nomeadas em apenas 3,30% dos casos. O baixo desempenho obtido pelos métodos de ponderação utilizando Entidades Nomeadas ocorreu principalmente por causa da baixa quantidade de entidades nomeadas extraídas por documento adotando a ferramenta CoreNLP.

Os métodos de ponderação baseados em grafos apresentaram, em sua grande maioria, resultados inferiores aos métodos estatísticos. Esses resultados reforçam que a centralidade das informações não desempenha um papel tão significativo quanto os aspectos de posição e frequência para a sumarização monodocumento. Esse mesmo comportamento também foi observado nos experimentos apresentados no Capítulo 3. Além disso, os métodos baseados em grafos incluem um custo computacional maior, devido ao processo de construção do grafo de conceitos.

Na sumarização monodocumento, especialmente em artigos de notícias, a posição das sentenças tem demonstrado ser um dos melhores métodos para a tarefa de pontuação de sentenças. Tal comportamento também foi observado nos experimentos discutidos neste apêndice. Trabalhos anteriores no contexto da sumarização multidocumento (GILLICK et al., 2009; CAO et al., 2015) obtiveram melhores resultados utilizando o método de frequência dos documentos do que adotando outros métodos para a ponderação dos conceitos. Nos experimentos realizados neste trabalho, no contexto da sumarização monodocumento de artigos de notícias, o método de posição de sentenças obteve resultados superiores do que os outros métodos baseados em frequência e centralidade. Tais resultados reforçam a importância da posição das informações para a sumarização monodocumento, e demonstram que o aspecto de posição das sentenças deve ser melhor explorado por abordagens baseadas em conceitos direcionadas para sumarização monodocumento.

Tabela 41 – Resultados (%) e desvio padrão entre parênteses das avaliações dos métodos de ponderação baseados em grafos no corpus CNN. O melhor desempenho global está destacado em negrito, e o grupo de configurações estatisticamente semelhantes, se existir, é indicado pelo símbolo †.

Métodos baseados em Grafo	Representações	CNN		
		IS	R-1	R-2
Betweenness	Unigrama	19,66	51,42 (18,17)	29,42 (23,58)
	Bigrama	17,96	47,51 (19,10)	26,35 (23,23)
	Entidade Nomeada	18,05	45,02 (18,39)	24,80 (22,09)
	Dep. Sintática G.	17,94	48,90 (19,08)	26,92 (23,52)
	Dep. Sintática R.	17,90	48,77 (19,22)	26,86 (23,74)
Closeness	Unigrama	18,36	50,52 (20,00)	28,40 (24,65)
	Bigrama	11,58	38,14 (18,01)	16,48 (19,59)
	Entidade Nomeada	17,40	44,79 (19,45)	24,42 (22,82)
	Dep. Sintática G.	11,90	39,00 (20,40)	18,16 (21,52)
	Dep. Sintática R.	11,65	38,52 (20,31)	17,73 (21,28)
Grau	Unigrama	20,11	53,26 (18,30)	30,88 (24,24)
	Bigrama	20,81	**53,85** (19,30)	**32,43** (25,21)
	Entidade Nomeada	18,79	46,45 (18,58)	26,07 (22,73)
	Dep. Sintática G.	19,69	53,55† (19,28)	31,34 (25,37)
	Dep. Sintática R.	19,59	53,42† (19,37)	31,21 (25,49)
Eigenvector	Unigrama	20,61	50,96 (19,09)	29,70 (23,79)
	Bigrama	17,62	44,69 (19,25)	24,70 (23,00)
	Entidade Nomeada	17,37	44,33 (18,54)	23,83 (22,01)
	Dep. Sintática G.	17,66	43,99 (19,19)	24,11 (22,68)
	Dep. Sintática R.	17,55	44,01 (19,15)	24,19 (22,77)
HITS Autoridade	Unigrama	19,30	49,91 (18,34)	28,11 (23,35)
	Bigrama	18,98	50,19 (19,67)	28,85 (24,65)
	Entidade Nomeada	18,57	45,90 (18,40)	25,51 (22,12)
	Dep. Sintática G.	18,95	50,81 (19,55)	29,09 (24,79)
	Dep. Sintática R.	18,83	50,73 (19,63)	28,90 (24,75)
HITS *Hub*	Unigrama	19,10	49,71 (18,49)	27,92 (23,15)
	Bigrama	19,51	50,49 (19,65)	29,42 (24,69)
	Entidade Nomeada	19,09	46,22 (18,79)	26,10 (22,85)
	Dep. Sintática G.	18,82	50,68 (19,40)	28,93 (24,48)
	Dep. Sintática R.	18,67	50,59 (19,39)	28,79 (24,41)
PageRank	Unigrama	19,92	53,19 (18,66)	30,84 (24,66)
	Bigrama	20,61	53,59† (19,59)	31,98 (25,79)
	Entidade Nomeada	19,13	46,69 (18,86)	26,58 (22,97)
	Dep. Sintática G.	19,21	53,07 (19,54)	30,69 (25,66)
	Dep. Sintática R.	19,26	52,94 (19,52)	30,49 (25,63)
TextRank	Unigrama	18,54	51,06 (18,51)	28,56 (24,25)
	Bigrama	20,54	53,41 (19,36)	31,93 (25,27)
	Entidade Nomeada	18,59	45,45 (18,74)	25,46 (22,41)
	Dep. Sintática G.	19,69	53,55† (19,28)	31,34 (25,37)
	Dep. Sintática R.	19,59	53,42† (19,37)	31,21 (25,49)

Tabela 42 – Resultados (%) e desvio padrão entre parênteses das avaliações dos métodos de ponderação estatísticos nos corpora do DUC 2001-2002. O melhor desempenho global em cada corpus é destacado em negrito, e o grupo de configurações estatisticamente semelhantes, se existir, é indicado pelo símbolo †.

Métodos Estatísticos	Representações	DUC 2001		DUC 2002	
		R-1	R-2	R-1	R-2
Freq. das Sentenças	Unigrama	41,07 (8,87)	14,58 (9,35)	44,71 (9,03)	18,19 (9,75)
	Bigrama	42,09 (8,68)	15,82 (9,60)	45,69 (8,81)	19,29 (9,58)
	Entidade Nomeada	39,35 (9,32)	14,13 (9,82)	43,15 (8,82)	17,36 (9,50)
	Dep. Sintática G.	41,98 (8,54)	15,34 (9,56)	44,62 (9,03)	18,12 (10,05)
	Dep. Sintática R.	41,86 (8,67)	15,25 (9,67)	44,88 (9,10)	18,54 (10,18)
Freq. do Conceito	Unigrama	41,01 (8,76)	14,75 (9,37)	44,45 (9,20)	18,15 (9,92)
	Bigrama	41,94 (8,68)	15,76 (9,57)	45,62 (8,73)	19,09 (9,59)
	Entidade Nomeada	39,40 (9,38)	14,23 (9,90)	43,17 (8,81)	17,40 (9,52)
	Dep. Sintática G.	42,05 (8,86)	15,53 (9,81)	44,90 (9,15)	18,54 (10,18)
	Dep. Sintática R.	41,71 (8,72)	15,21 (9,68)	44,80 (9,06)	18,56 (10,09)
FC-FIC	Unigrama	39,60 (8,97)	13,79 (9,29)	43,51 (9,23)	17,48 (9,78)
	Bigrama	41,81 (8,86)	15,67 (9,74)	45,29 (8,84)	18,89 (9,70)
	Entidade Nomeada	38,96 (9,58)	13,84 (10,07)	42,98 (8,97)	17,25 (9,67)
	Dep. Sintática G.	41,25 (8,30)	14,95 (9,09)	44,57 (8,85)	18,16 (9,83)
	Dep. Sintática R.	41,30 (8,40)	14,97 (9,17)	44,41 (8,86)	18,04 (9,83)
Pos. do Conceito	Unigrama	40,44 (9,11)	14,11 (9,45)	44,18 (9,23)	17,91 (9,88)
	Bigrama	41,40 (8,49)	15,05 (9,13)	45,55 (8,89)	19,17 (9,73)
	Entidade Nomeada	39,25 (9,21)	13,99 (9,86)	42,84 (8,86)	17,07 (9,45)
	Dep. Sintática G.	41,10 (8,66)	14,73 (9,66)	44,24 (9,15)	18,03 (9,92)
	Dep. Sintática R.	40,99 (8,69)	14,47 (9,51)	44,22 (9,09)	18,03 (9,93)
Pos. da Sentença	Unigrama	38,65 (9,56)	13,97 (9,52)	43,07 (9,53)	17,37 (10,05)
	Bigrama	43,02† (9,38)	17,50† (10,63)	**47,31** (8,86)	**21,59** (10,16)
	Entidade Nomeada	40,38 (9,69)	15,47 (10,20)	44,29 (9,14)	18,84 (9,94)
	Dep. Sintática G.	43,47† (9,35)	18,01† (10,32)	47,05† (9,06)	21,39† (10,41)
	Dep. Sintática R.	**43,49** (9,54)	**18,21** (10,40)	47,11† (9,07)	21,37† (10,39)

Tabela 43 – Resultados (%) e desvio padrão entre parênteses das avaliações dos métodos de ponderação baseados em grafos nos corpora do DUC 2001-2002. O melhor desempenho global em cada corpus é destacado em negrito, e o grupo de configurações estatisticamente semelhantes, se existir, é indicado pelo símbolo †.

Métodos baseados em Grafos	Representações	DUC 2001		DUC 2002	
		R-1	R-2	R-1	R-2
Betweenness	Unigrama	39,98 (8,79)	13,82 (9,51)	43,53 (8,98)	17,33 (9,77)
	Bigrama	39,86 (8,11)	13,52 (8,44)	42,76 (9,06)	16,75 (9,74)
	Entidade Nomeada	38,94 (9,49)	13,88 (9,78)	42,64 (8,92)	16,81 (9,43)
	Dep, Sintática SR	39,48 (8,15)	13,05 (8,50)	43,28 (9,22)	17,00 (9,86)
	Dep, Sintática R	39,76 (8,27)	13,44 (8,81)	43,03 (9,24)	16,92 (9,93)
Closeness	Unigrama	38,38 (9,36)	12,84 (9,42)	42,59 (9,89)	16,61 (10,20)
	Bigrama	37,58 (9,42)	12,18 (9,41)	42,01 (9,70)	15,79 (10,10)
	Entidade Nomeada	38,65 (9,68)	13,77 (9,99)	42,34 (9,16)	16,64 (9,61)
	Dep, Sintática SR	38,38 (9,58)	12,74 (9,47)	42,46 (9,54)	16,39 (10,15)
	Dep, Sintática R	38,33 (9,63)	12,79 (9,52)	42,13 (9,66)	16,05 (10,22)
Grau	Unigrama	40,50 (8,55)	14,17 (9,05)	44,13 (9,26)	17,64 (9,86)
	Bigrama	41,47† (8,85)	**15,23** (9,98)	**45,20** (8,62)	**18,81** (9,49)
	Entidade Nomeada	39,28 (9,73)	14,09 (10,14)	42,99 (8,79)	17,10 (9,45)
	Dep. Sintática G.	**41,84** (8,02)	15,01† (8,94)	44,61 (8,76)	18,03 (9,84)
	Dep. Sintática R.	41,67† (8,09)	14,94 (8,99)	44,52 (8,78)	18,11† (9,87)
Eigenvector	Unigrama	40,96† (8,79)	14,19 (9,31)	44,24 (8,81)	17,86 (9,63)
	Bigrama	39,89 (9,13)	14,20 (9,60)	42,69 (9,15)	17,16 (9,94)
	Entidade Nomeada	39,96 (9,24)	14,31 (9,93)	42,75 (8,78)	16,72 (9,41)
	Dep. Sintática G.	39,15 (9,22)	13,66 (9,53)	43,17 (9,02)	17,47 (9,83)
	Dep. Sintática R.	39,08 (9,22)	13,60 (9,53)	42,83 (8,93)	17,14 (9,67)
HITS Autoridade	Unigrama	39,89 (8,99)	13,70 (9,53)	43,36 (9,08)	16,97 (9,35)
	Bigrama	41,02† (9,23)	14,74† (9,59)	44,47 (9,08)	18,38† (9,86)
	Entidade Nomeada	39,62 (9,19)	13,92 (9,87)	42,55 (8,62)	16,56 (9,21)
	Dep. Sintática G.	41,04† (8,93)	14,86† (9,58)	44,03 (8,71)	17,85 (9,67)
	Dep. Sintática R.	41,06† (8,56)	14,88† (9,42)	43,72 (8,84)	17,56 (9,76)
HITS *Hub*	Unigrama	40,42 (8,93)	14,13 (9,67)	43,55 (9,58)	17,04 (10,16)
	Bigrama	41,13† (9,21)	15,05† (9,96)	44,80† (9,39)	18,79† (10,14)
	Entidade Nomeada	39,85 (9,20)	14,17 (10,03)	43,13 (8,98)	17,25 (9,61)
	Dep. Sintática G.	40,79† (8,83)	14,54† (9,52)	44,28 (9,19)	18,18† (10,09)
	Dep. Sintática R.	40,57† (8,76)	14,43† (9,56)	44,12 (9,27)	18,07 (10,04)
PageRank	Unigrama	39,70 (9,19)	13,84 (9,40)	42,92 (9,37)	17,00 (9,81)
	Bigrama	41,12† (8,76)	14,92† (9,49)	44,19 (9,08)	17,87 (9,83)
	Entidade Nomeada	39,04 (9,59)	13,84 (9,95)	42,21 (8,98)	16,47 (9,57)
	Dep. Sintática G.	41,10† (8,68)	14,73† (9,29)	43,85 (8,98)	17,55 (9,85)
	Dep. Sintática R.	41,25† (8,36)	14,78† (8,93)	43,63 (9,02)	17,40 (9,77)
TextRank	Unigrama	39,19 (9,63)	13,76 (9,69)	43,55 (9,58)	17,04 (10,16)
	Bigrama	41,40† (9,28)	15,10† (10,41)	44,80† (9,39)	18,79† (10,14)
	Entidade Nomeada	38,64 (9,65)	13,61 (9,88)	42,32 (9,08)	16,66 (9,68)
	Dep. Sintática G.	**41,84** (8,02)	15,01† (8,94)	44,61 (8,76)	18,03 (9,84)
	Dep. Sintática R.	41,67† (8,09)	14,94† (8,99)	44,52 (8,78)	18,11† (9,87)

A.3 Considerações Finais do Apêndice

Neste apêndice foram apresentados os experimentos conduzidos para avaliar treze métodos para a ponderação da importância, e cinco formas de representação para os conceitos em uma abordagem baseada em conceitos para sumarização monodocumento. Os resultados empíricos demonstram as seguintes conclusões: **(i)** Uma clara superioridade da adoção de bigramas como conceitos em documentos de notícias. Esses resultados corroboram com os resultados obtidos por Gillick et al. (2009) e Schluter e Søgaard (2015) para a sumarização multidocumento; **(ii)** De modo global, o método de posição de sentenças obteve um desempenho estatisticamente superior a todos os outros métodos de ponderação de conceitos; e **(iii)** Entre os métodos de ponderação baseados em grafos, utilizar o grau dos vértices como pontuação produziu os melhores resultados nas medidas avaliadas.

As conclusões obtidas neste Apêndice dão suporte a escolha de bigramas como representação de conceitos pelo método proposto no Capítulo 4. Além disso, o bom desempenho obtido pelo método de posição de sentenças observado nos experimentos realizados, ressaltou a necessidade de integrar essa informação de forma mais explícita nas abordagens baseadas em conceitos para sumarização monodocumento. Por fim, os bons resultados obtidos pelos métodos de frequência e posição das sentenças, e a alta diversidade entre os resumos gerados por eles, despertou a possibilidade da combiná-los para melhor estimar a relevância dos conceitos extraídos.