Pós-Graduação em Ciência da Computação

Jamilson Batista Antunes

Uma Abordagem para Sumarização Automática Semi-Extrativa

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

Recife
2018

Jamilson Batista Antunes

Uma Abordagem para Sumarização Automática Semi-Extrativa

Tese apresentada ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Doutor em Ciência da Computação.

Área de Concentração: Inteligência Artificial
Orientador: Prof. Dr. Rafael Dueire Lins

Recife
2018

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

A636a Antunes, Jamilson Batista
Uma abordagem para sumarização automática semi-extrativa / Jamilson Batista Antunes. – 2018.
175 f.: il., fig., tab.

Orientador: Rafael Dueire Lins.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2018.
Inclui referências.

1. Inteligência artificial. 2. Sumarização. I. Lins, Rafael Dueire (orientador). II. Título.

006.31 CDD (23. ed.) UFPE- MEI 2018-141

Jamilson Batista Antunes

Uma Abordagem para Sumarização Automática Semi-Extrativa

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

Aprovado em: 12/11/2018.

_____

Orientador: Prof. Dr. Rafael Dueire Lins

BANCA EXAMINADORA

_____

Prof. Dr. Frederico Luiz Gonçalves de Freitas
Centro de Informática/ UFPE

_____

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

_____

Prof. Dr. Steven John Simske
Departamento de Sistemas e Engenharia Mecânica /
Universidade do Estado do Colorado

_____

Prof. Dr. Rafael Ferreira Leite de Mello
Departamento de Computação / UFRPE

_____

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

Decido este trabalho a minha família e minha esposa que foram porto seguro perante as dificuldades durante este percurso.

AGRADECIMENTOS

Agradeço à minha família, especialmente meus pais (Jose Alcides Antunes e Elza Maria do Nascimento Antunes) e meus irmãos e irmãs (Josué Batista Antunes, Jardel Batista Antunes, Ester Mirian Batista Antunes e Sthefani Batista Antunes), por sempre me apoiarem incondicionalmente e por me incentivarem a nunca desistir dos meus sonhos.

Á minha amada esposa, Rafaelle Mayanne Silva Salgado Cavalcanti, que selamos nossa união no penúltimo desse doutorado, por todo apoio, compreensão e amor cedidos durante todos os momentos do meu doutorado.

Ao meu orientador, professor Dr. Rafael Dueire Lins, pela amizade, confiança, palavras de incentivo, dedicação e orientação.

Aos membros do grupo de pesquisa em Sumarização Automática de Textos do CIN-UFPE, pelas contribuições e grandes ensinamentos.

A todos os meus amigos, em especial João Emanoel, Hilário Tomaz, Renê Gadelha, Alex Nery, Francisco Airton e Adriano Ferraz por todos os momentos, cujos lembrarei por toda a minha vida.

Aos velhos amigos da Fitec e novos da Thoughtworks, pela convivência durante os quatro anos de doutorado e pelas conversas e apoio nas minhas horas de ausência do trabalho para reuniões com meu orientador.

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, a HP Brasil, HP Labs e ao Steven J. Simsked por confiar no nosso grupo de pesquisa para desenvolvedor pesquisas na área de Sumarização Automática de Textos.

RESUMO

A Sumarização Automática de Textos (SAT) consiste em criar versões comprimidas de um ou mais documentos de texto, mantendo as informações essenciais dos documentos. Essa área de pesquisa vem se tornando cada vez mais importante, já que potencialmente auxilia o processamento de grandes volumes de informações, permitindo destacar as informações mais relevantes para o usuário. Além de poder reduzir significativamente a quantidade de tempo que as pessoas despendem em tarefas de leitura. O uso de Processamento de Linguagem Natural (PLN) revela-se benéfico ao processo de sumarização, principalmente quando se processam textos sem nenhuma estrutura e/ou padrão definido. Dentre as variações do processo de sumarização, as técnicas extrativas são as mais bem estudadas até o momento. O principal foco da investigação mais recente sobre sumarização extrativa é a otimização de algoritmos que visam obter o conteúdo relevante expresso nos textos originais. Porém, os ganhos relacionados com o aumento da complexidade desses algoritmos não foram ainda comprovados, já que os sumários continuam a ser difíceis de ler. 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 estratégias atuais de sumarização preocupam-se principalmente em maximar a informatividade dos resumos, sem levar em consideração a qualidade textual. Investigações recentes na literatura e experimentos conduzidos neste trabalho demonstram que essas características são uma limitação significativa, já que os resumos devem ser gerados serem lidos por seres humanos. Nesse contexto, a presente tese propõe uma abordagem para sumarização automática semi-extrativa que compreende à resolução de anáforas pronominais, reinserção de pronomes e redução de sentenças. Além disso, avaliaram-se medidas para estimar a qualidade textual de resumos candidatos sem o uso de um resumo de referência. Esta tese foca a sumarização automática numa perspectiva diferente, estudando o impacto da sumarização extrativa na abstrativa, a fim de produzir um sumário de melhor qualidade textual em termos de informatividade, legibilidade, fluência e coesão. Diversos experimentos foram conduzidos nos principais corpora da área, visando avaliar diferentes aspectos das abordagens propostas nas tarefas de sumarização monodocumento. Os resultados obtidos demonstram que as soluções apresentadas são capazes de aumentar a qualidade textual e a informatividade dos resumos gerados, com base nas avaliações humanas e automáticas para diversos sistemas do estado da arte.

Palavras-chaves: Inteligência Artificial. Sumarização Automática de Textos. Sumarização Extrativa Monodocumento. Análise de Informatividade, Coesão e Legibilidade de Textos.

ABSTRACT

Automatic Text Summarization (ATS) consists of creating compressed versions of one or more text documents, while retaining the essential document information. This research area is becoming increasingly more important, since it can potentially help processing large volumes of data, allowing the most relevant information to be highlighted to the user. In addition to this, ATS will be able to significantly reduce the amount of time people spend on reading. The use of Natural Language Processing (NLP) has proven to be advantageous to the summarization process, especially when processing texts with no defined structure and/or pattern. Among the variations of the summarization process, the extractive techniques are the best studied so far. The main focus of the most recent research on extractive summarization is the optimization of algorithms aimed at obtaining the relevant content expressed in the original texts. However, the gains associated with increasing the complexity of those algorithms have not yet been assessed, since the summaries are still difficult to read. Despite the advances made in recent years, there is still a big difference between automatically generated summaries and those written by humans. Most of the current summarization strategies are mainly concerned with maximizing the informativeness of summary, disregarding the text quality. Recent investigations in the literature and experiments conducted in this work demonstrate that those features yield a significant limitation, since the abstracts generated being must be read by humans. In such a context, this thesis proposes an approach for semi-extractive automatic summarization in which, it includes the resolution of pronominal anaphoras, the reinsertion of pronouns to increase the readeability of the text, and the reduction of the size of sentences, allowing to increase the informativeness of the generated summary with the same number of words. Besides all that, we evaluated several measures present in the literature to estimate the quality of abstracts without using a reference summary. This thesis addresses the problem of automatic summarization in a different perspective, studying the impact of extractive summarization on the abstract, in order to produce the best possible summary in terms of informativeness, readability, fluency and cohesion. Several experiments were conducted in the main corpora of the area, aiming to evaluate different aspects of the proposed approaches in the tasks of single-document summarization. The results obtained show that the proposed solutions are able to increase the textual quality and the informativeness of the abstracts generated, based on human and automatic evaluations for the different state of the art systems.

Keywords: Artificial intelligence. Automatic Text Summarization. Extractive Single-document Summarization. Analysis of Informativity, Cohesion and Text Readability.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

AES	*Anaphoric Expression Solver*
AESC	Resolução de Expressões Anafóricas no Corpus
AESS	Resolução de Expressões Anafóricas no Sumário
API	*Application program interface*
ATS	*automatic text summarization*
BNC	*British National Corpus*
CoreNLP	*Stanford Natural Language Processing Toolkit*
CS	similaridade do cosseno
GATE	*General Architecture for Text Engineering*
HIT	*Human Intelligence Task*
HITL	*Human-in-the-loop*
KL	*Kullback-Lieber*
LRs	Recursos de Linguagem
MRE	entidade mais representativa
NER	entidades nomeadas
NLP	*Natural Language Processing*
PLN	Processamento de Linguagem Natural
PRs	Recursos de Processamento
RC	Resolução de Correferência
SA	Sumarização Automática
SAT	sumarização automática de texto
SCRS	Sistema de Resolução de Correferência de Stanford
SDKs	*Software Development Kits*
SSF	Fluxo Padrão de Sumarização
TF-IDF	*Term Frequency-Inverse Document Frequency*
VRs	recursos de visualização

SUMÁRIO

1 INTRODUÇÃO

O rápido crescimento da Internet resultou num grande aumento da quantidade de dados disponível, especialmente em relação aos documentos de textos (artigos de notícias, livros eletrônicos, artigos científicos, *blogs*, *microblogs*, redes sociais, etc.). O processamento desses dados para a extração de informação tem se tornado cada vez mais difícil, tanto para humanos quanto para máquinas. Enquanto humanos não têm a capacidade nem o tempo de ler e compreender todas as informações disponíveis de seu interesse, as máquinas perdem em precisão e desempenho ao lidar com tamanha quantidade e variedade de documentos. O estudo realizado por Bosch, Bogers e Kunder (2016) mostra que nos últimos nove anos estima-se, a partir dos resultados obtidos pelos dois principais motores de busca (Google e Bing), que existam aproximadamente 50 bilhões de páginas de informações na Web. Nesse cenário, a Sumarização Automática (SA) mostra-se como uma tarefa que pode auxiliar significativamente no fornecimento de informações para simplificar a leitura das pessoas. A tarefa de SA consiste na produção automática de uma versão mais curta de um ou mais textos-fonte, chamada de sumário ou resumo (HAHN; MANI, 2000). O sumário produzido a partir de um único texto fonte é denominado sumário *monodocumento*. Já a SA *multidocumento* consiste na produção de um único sumário a partir de um conjunto de textos-fonte/documentos sobre um mesmo assunto (ZAJIC; DORR; LIN, 2008; FATTAH; REN, 2009).

Segundo Torres-Moreno (2014) a sumarização humana de texto é um processo cognitivo e requer um completo entendimento do documento ou fato a ser sintetizado. Esse processo não é exato, ou seja, se diferentes pessoas sumarizarem um mesmo documento, é provável que resumos diversos sejam gerados.

No geral, sumarizar (resumir) conteúdos é uma tarefa comum no cotidiano de qualquer pessoa. As informações disponibilizadas em diferentes formatos: texto, vídeo, áudio, etc. são constantemente resumidas, isso mostra que um resumo é um instrumento de comunicação e atualização em áreas como meio acadêmico, negócios e social. Por exemplo, quando numa conversa uma pessoa pergunta como foi um filme ou o que foi discutido na última reunião de negócios. Outro exemplo são os resumos de documentos, tais como teses, dissertações, artigos ou textos de notícias. Esses resumos auxiliam o leitor no entendimento de um documento de forma rápida e facilita uma tomada de decisão, isto é, se o documento é relevante para uma leitura completa. Para isso, o resumo deverá ser informativo, legível e coeso para facilitar a compreensão por parte do leitor.

O foco da sumarização textual é distinguir no texto original o conteúdo relevante daquilo que poderá ser descartado, para a composição de uma versão resumida do texto fonte. Segundo Mani (2001), quando uma pessoa resume um texto, tenta primeiro identificar a sua essência, e em seguida adiciona gradualmente informação retirada do texto original

para complementar o sumário. Entretanto, esse processo está sujeito a diversos fatores influenciadores: perfil de quem está escrevendo o sumário, perfil dos potenciais leitores, e importância subjetiva que o autor e o leitor possam atribuir ao teor do texto, fazendo com que não só o conteúdo, mas também a estrutura do sumário possa estar sujeita à forma na qual se pretende representar o conteúdo do texto original.

Baseado nos comportamentos descritos acima, verifica-se o quão complexa e abstrata é a tarefa de sumarização, e quão desafiadora é a ideia de automatizá-la. Nenkova e McKeown (2012) definem que a sumarização automática de texto (SAT) é um processo automático que visa gerar uma versão mais compacta (resumo) de um ou mais textos fonte, mantendo suas informações mais relevantes. A sumarização automática de texto, do inglês *automatic text summarization* (ATS), visa criar uma versão comprimida (resumo) de um ou mais documentos e extrair as informações essenciais a partir deles (WANG et al., 2010). O objetivo principal de um resumo é apresentar as principais ideias de um documento em menos espaço (RADEV; HOVY; MCKEOWN, 2002). Ao definir um resumo, Hovy (2004) e Radev, Hovy e McKeown (2002) concordam que este deve ser no máximo metade do texto original.

A Sumarização Automática pode ser classificada em *extrativa* e *abstrativa* (LLORET; PALOMAR, 2012). Um sistema extrativo seleciona um conjunto de frases mais significativas de um documento, exatamente como elas aparecem, para formar o resumo. Um sistema abstrativo produz um resumo que inclui palavras e frases diferentes das que ocorrem no documento de origem. Portanto, um sumário abstrativo consiste de ideias ou conceitos retirados do documento original, mas são reinterpretados e apresentados de uma forma diferente.

Os resumos gerados podem ser de dois tipos: *baseados em consultas (query)* ou *genéricos* (GONG; LIU, 2001; DUNLAVY et al., 2007; OUYANG et al., 2011). Para o primeiro tipo, o conteúdo do sumário é relacionado com a *query* usada na consulta, enquanto um sumário genérico apresenta, num sentido geral, as informações mais relevantes de um documento. Para gerar os resumos, geralmente, são adotadas duas abordagens mais utilizadas na literatura: *supervisionada* e *não-supervisionada* (MANI, 1999; FATTAH; REN, 2009; RIEDHAMMER; FAVRE; HAKKANI-TüR, 2010). Na abordagem supervisionada é necessário, para as técnicas de aprendizagem, um conjunto de dados rotulados ou anotados de treinamento para selecionar o conteúdo importante dos documentos. Por outro lado, os sistemas baseados na abordagem não-supervisionada não necessitam de dados de treinamento, eles aplicam heurísticas para extrair informações altamente relevantes e gerar um resumo (FATTAH; REN, 2009). Por isso, a última abordagem é mais indicada para sumarização genérica.

A sumarização automática vem sendo explorada desde a década de 1950, quando surgiram os primeiros métodos para a produção de extratos[1], sendo o método de palavras-chave (LUHN, 1958) o mais significativo até então. Entretanto, como métodos "cegos"

[1] Referem-se à palavras, frases, sentenças ou parágrafos extraídos do texto fonte.

(sem uma profunda análise do texto), fazendo uso de técnicas de frequência de palavras, os resultados apresentavam inúmeros problemas, quer de coesão[2], quer de coerência (HUTCHINS, 1987), razão pela qual a área ficou praticamente estagnada nas décadas seguintes, voltando a ser objeto de interesse com o advento da Internet e, portanto, com o aumento considerável de documentos disponíveis *online* e com a necessidade de se "digerir" informações em larga escala, isto é, em grande quantidade e no menor tempo possível.

Jones (1993) atribui a ausência de progresso significativo na área até meados da década de 1990 à dificuldade de se modelar adequadamente o processo, resultando na impossibilidade de se obter sumários automáticos de qualidade. Contudo, ela observa que, para aplicações restritas, é possível explorar critérios estruturais ou composicionais em certa profundidade, baseados nos aspectos linguísticos do texto fonte.

A Sumarização Monodocumento Extrativa foi bastante explorada e discutida por vários autores (por exemplo, Luhn (1958), Edmundson (1969), O'Donnell (1997), Marcu (2000), Pardo, Rino e Nunes (2003), Yeh et al. (2005), Fattah e Ren (2009), Alguliev et al. (2011), Ferreira et al. (2013), Mendoza et al. (2014), Cabral et al. (2015)). Segundo Gambhir e Gupta (2016), a comunidade está se concentrando mais em sumarização extrativa, tentando alcançar resumos mais coerentes e significativos. Apesar de tentarem resolver o problema da falta de coesão textual (que implicará, quase que certamente, na falta de coerência), as alternativas apontadas acima eventualmente levam a resumos mais longos do que o desejado, ou seja, é necessário incluir um número maior de sentenças para que o sumário faça sentido. Além disso, ainda não garantem a coesão, tampouco a coerência. Uma característica fundamental que os resumos devem ter para que sejam úteis para todas essas aplicações é que eles mantenham as informações mais relevantes dos documentos originais para atender aos requisitos dos usuários. Além disso, como os resumos gerados são destinados à leitura humana, eles devem ser fluentes, ou seja, coesos, coerentes, legíveis e gramaticalmente corretos.

Para exemplificar o foi discutido até aqui, o quadro a seguir mostra dois sumários, o primeiro (I) não apresenta problema de coesão, a mensagem está clara, objetiva e legível. No entanto, o segundo (II) sumário apresenta problemas de legibilidade e coesão, ou seja, as sentenças são vagas e contêm pronomes não conectados à um substantivo, dificultando o entendimento da informação veiculada na notícia[3].

[2] Coesão está relacionada com mecanismos linguísticos do texto, que são responsáveis por estabelecer uma conexão de ideias. A coesão cria relações entre as partes do texto de modo a guiar o leitor relativamente a uma sequência de fatos.

[3] https://edition.cnn.com/2018/12/11/business/huawei-apple-china-tech-us/index.html

(I) Arianna Huffington describes herself as a “sleep evangelist,” has nap rooms in her offices at the AOL headquarters in New York and tries to start every day with meditation. Huffington, 62, founded Huffington Post in 2005, and two years ago sold it to AOL for $315 million.

(II) The notice has since been taken down and the company did not respond to CNN’s attempts to contact them. "We don’t have guns or cannons, we as common citizens only have freedom of speech,"she told CNN. The chamber claims to have as many as 500 members. "The stakes are very, very high in terms of the overall trade dispute between the US and China,"he said.

O problema de se construir sumários automaticamente remete às máximas de Grice (1975), a saber:

- Qualidade: informar precisamente somente o que pode ser evidenciado no texto fonte;
- Quantidade: se ater ao número máximo de palavras e/ou sentenças de forma objetiva e sem redudância;
- Relevância: transmitir somente o necessário e solicitado, dependendo da meta comunicativa e do conhecimento do leitor;
- Modo: evitar obscuridade e ambiguidade e escrever de forma coerente e breve.

Na sumarização automática, a máxima de qualidade pode ser expandida aos critérios de qualidade linguística considerados pelas conferências DUC[4] (*Document Understanding Conference*) e TAC[5] (*Text Analysis Conference*): gramaticalidade, não redundância, clareza referencial, foco e estrutura, e coerência (DANG, 2005).

Segundo McNamara et al. (2014), há evidências consideráveis de que a falta de coesão determina de forma crítica o quão desafiador é um texto para SAT e como o leitor vai entendê-lo. Mesmo para modalidade mais madura de sumarização de notícias, os sistemas tornaram-se bons em selecionar conteúdo importante, mas a qualidade linguística dos resumos gerados é bastante fraca.

Kaspersson et al. (2012), Rennes e Jonsson (2014), fizeram estudos sobre os erros mais comuns em sumarização extrativa, eles apontaram a ausência de coesão ou contexto e expressões anafóricas quebradas[6] como a principal causa dos erros. Além desses autores, Gonçalves, Rino e Vieira (2008), Smith, Henrik e Arne (2012) também haviam apontado problemas de legibilidade nos sumários gerados. Para contornar esses problemas, Silveira (2015) propôs um conjunto de tarefas para melhorar a coesão de um sumário extrativo após ele ter sido gerado.

[4] http://duc.nist.gov/pubs.html
[5] http://www.nist.gov/tac/about/
[6] Pronomes livres e não contextualizados, sem conexão a um substantivo no sumário

Além dos desafios para gerar sumários com qualidade, outro problema relevante é como avaliar esses sumários. Os autores em sua maioria avaliam os sumários utilizando medidas intrínsecas baseadas em estatísticas como o ROUGE (LIN, 2004). No entanto, essas medidas não levam em consideração a qualidade textual ou do conteúdo. Segundo Louis (2013) medidas de qualidade de texto podem ajudar os sistemas de sumarização à criar textos coesos e legíveis.

Neste trabalho, assim como em Pitler e Nenkova (2008), Louis (2013), McNamara et al. (2014), supõe-se que a coesão possa ser medida e que a coerência está relacionado ao processo de interpretação do leitor, ao passo que a coesão reside no texto ou discurso (CARRELL, 1982; GIVÓN, 1995). Uma premissa importante é que os elementos textuais (por exemplo, correferências pronominais), os quais influenciam na coesão, podem ser medidos e quantificados diretamente. Coerência, por outro lado, refere-se à forma como o leitor compreende um texto ou discurso e, portanto, a coerência do texto só pode ser medida de forma indireta. Pode-se fazer isso, por exemplo, fazendo perguntas ao leitor, apresentando tarefas que investigam a compreensão e avaliação da memória para a informação veiculada no texto. A coerência de uma representação mental surge como uma função do número de associações ou conexões construídas pelo leitor. Quanto maior o número de conexões entre as ideias, mais coerente o texto.

No decorrer deste trabalho, será utilizado, em diversos momentos, o termo "coesão". Mas, neste ponto, é importante ressaltar que a coesão é um termo genérico, referindo-se aos muitos elementos diferentes no texto que contribuem coletivamente para a coesão. Quando um elemento contribui para coesão, segundo McNamara et al. (2014) o termo "*cohesive cue*" (sugestão coesiva) é o mais indicado. Assim, por exemplo, resolução de anáforas pronominais e a inserção de pronomes são potenciais sugestões coesivas. Portanto, nesse trabalho pretende-se, após um estudo minucioso desses elementos, apresentar abordagens que possibilitem a geração de sumários semi-extrativos com alta qualidade textual.

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

1.1 CARACTERIZAÇÃO DA ÁREA

A tarefa de sumarização automática refere-se a gerar um resumo contendo as informações mais relevantes, a partir de um conjunto de documentos de entrada. Os resumos produzidos por um sistema de sumarização podem ser classificados em função dos diversos fatores que influenciam a sua criação. 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, dentre outras (LLORET; PALOMAR, 2012).

Os sumários podem ser classicados com base numa grande diversidade de critérios (JONES, 1998; HOVY; LIN, 1999), para dar uma visão das diferentes classicações, mais recorrentes na literatura, que tentam classicar um sumário, a tabela 2 apresenta uma classicação que permite a caracterização de sumários produzidos por sistemas de sumarização automática por uma vasta gama de propriedades.

Tabela 2 – Classificação das abordagens de SAT baseada em diversas características.

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

Com base no número de documentos, a tarefa de sumarização pode ser classificada em duas categorias: monodocumento e multidocumento (ZAJIC et al., 2007; FATTAH; REN, 2009):

- Na sumarização monodocumento, um resumo é gerado a partir de um único documento de entrada;
- Na sumarização multidocumento muitos documentos relacionados com um mesmo evento ou assunto são usados para gerar um único resumo. Considera-se que a sumarização monodocumento pode ser estendida para gerar sumarização de múltiplos documentos. 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 os tópicos essenciais do texto fonte, pretendendo informar sobre o teor do texto fonte sem transmitir conteúdo específico como detalhes de resultados, argumentações ou conclusões. Geralmente esse tipo de resumo é parecido como uma tabela de conteúdo, listando os tópicos mais relevantes;

- Os resumos informativos têm por objetivo apresentar as informações mais relevantes do(s) documento(s) de entrada, ou seja, ser uma versão mais compacta, e, em alguns casos, eles podem substituir a leitura dos documentos originais. O sumário deve explicar determinado conceito com o maior nível de detalhe possível, para determinada taxa de compressão;

- Os resumos críticos ou avaliativos apresentam, de forma sucinta, avaliações, comparações e opiniões sobre produtos, serviços, ou outros tópicos dos documentos originais. Esse tipo de resumo também está associado à polaridade (positiva, negativa ou neutra) das avaliações de cada assunto discutido nos textos de entrada. Portanto, sumários avaliativos servem de revisões críticas.

Em relação ao idioma, os sistemas de sumarização automática podem ser de três tipos (LLORET; PALOMAR, 2012): monolíngue, multilíngue e idiomas cruzados.

- A sumarização monolíngue gera sumários cujo conjunto de textos fonte se encontra todo no mesmo idioma. Esse tipo de sumarização é a mais comum na literatura;

- A sumarização multilíngue ocorre quando o(s) documento(s) de entrada estão em vários idiomas, por exemplo, inglês, português ou espanhol, 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.

Os resumos podem ser classificados de acordo seu propósito (SAGGION; POIBEAU, 2013):

- Sumário Genérico: o processo de sumarização é executado sem levar em consideração quaisquer informações que o usuário necessita, em outras palavras, todos os tópicos do texto original são igualmente importantes e devem ser incluídos no sumário;

- Sumário Baseado em Consultas: o processo de sumarização é realizado considerando as informações necessárias para responder uma determinada consulta do usuário;

- Sumário 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

- Sumário baseado 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 de 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 e copiam as sentenças mais relevantes dos documentos de entrada e as utilizam sem nenhuma alteração para compor o resumo. Este tipo de sumário está habitualmente associado a sistemas de sumarização que adotem uma abordagem superficial;

- Nas abordagens Abstrativas o texto resumido é uma interpretação do texto original. O processo de produção envolve reescrever o texto original numa versão mais curta, substituindo conceitos extensos por outros análogos mais curtos. Um sistema que produza sumários abstrativos analisa o texto fonte e tenta apresentar a sua compreensão do texto de uma forma humanamente inteligível, ou seja, em linguagem natural clara. Para isso, utiliza operações como compressão de sentenças (ZAJIC et al., 2007), fusão de sentenças (FILIPPOVA, 2010) e geração de linguagem natural (KHAN; SALIM; KUMAR, 2015).

Além das classificações apresentadas anteriormente, as abordagens de sumarização também podem ser classificadas de acordo com o gênero dos documentos de entrada (LLORET; PALOMAR, 2012):

- Sumarização de Notícias: resumos de artigos de notícias;

- Sumarização Especializada: resumos de documentos especializados em um único domínio, por exemplo, Ciência, Medicina, Biologia, entre outros;

- Sumarização Literária: resumos de documentos narrativos, textos literários, entre outros;

- Sumarização de Enciclopédias: sumários de documentos de enciclopédias, por exemplo, a Wikipédia;

- Sumarização em Redes Sociais: resumos de postagens em redes sociais;

- Sumarização de atas de reuniões: resumos de textos produzidos em uma ou mais reuniões (BANERJEE; MITRA; SUGIYAMA, 2015); entre outros tipos de documentos.

1.2 CONTEXTO DESTA PESQUISA

Dados os problemas de coesão e legibilidade que permeiam a tarefa sumarização automática de texto. Neste trabalho propõe-se uma abordagem para sumarização semi-extrativa que faz o uso de tarefas de resolução de anáforas, geração de pronomes e redução de sentenças para melhorar a legibilidade, coesão e fluência dos resumos. Portanto, possibilitando ao leitor maior compreensão do conteúdo, desprendendo de esforço para o entendimento da informação vinculada.

Segundo Mani (2001), existem duas abordagens tradicionais para a SA em geral: a superficial (ou empírico/estatística) e a profunda (ou fundamental). A primeira faz uso de pouco conhecimento linguístico (nível léxico e morfossintático), produzindo sumários formados por extratos do texto por meio de frequência de palavras (ou termos), sendo suas vantagens a escalabilidade e a robustez. A segunda faz uso de mais conhecimentos linguísticos (tais como regras gramaticais, semântica, resolução de anáfora, conhecimento discursivo, etc.), atingindo o nível semântico e discursivo, produzindo melhores resultados, entretanto, com maior custo.

Apesar do desenvolvimento relativamente simples e de baixo custo, é consenso na área que os métodos superficiais[7] comumente produzem sumários de qualidade inferior aos sumários produzidos por métodos profundos[8] (MANI, 2001; JONES, 2007). Diante do exposto, (SILVEIRA; BRANCO, 2012; SILVEIRA, 2015) mostrou em seus experimentos que abordagens profundas que usam a edição de um sumário extrativo, gerando assim um novo sumário, é uma alternativa para melhorar a qualidade textual dos resumos gerados por técnicas e sistemas do estado da arte (NENKOVA; MCKEOWN, 2012; FERREIRA et al., 2013). 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.

Este trabalho foca em uma abordagem profunda para sumarização automática mono-documento semi-extrativa.

[7] Tais métodos caracterizam-se pelo tratamento mais simples dos fenômenos linguísticos e pelo baixo custo na produção dos sumários.

[8] Os métodos profundos caracterizam-se pela utilização de conhecimento linguístico de nível morfológico, sintático, semântico e até pragmático-discursivo na tarefa de seleção de conteúdo para construir os sumários.

1.3 SOLUÇÕES PROPOSTAS

Propõe-se neste trabalho uma abordagem composta pelo desenvolvimento de estratégias para geração de sumários semi-extrativos independente de sumarizador; com a responsabilidade de fornecer:

1. **Tarefa de compressão de sumários**, de modo que, as sentenças dos resumos sejam reduzidas sem perder a informatividade e legibilidade, e que o sumarizador consiga incluir o maior número de informações nos sumários, e;

2. **Tarefas de fluência**, de maneira que, os sumários se tornem mais coesos e legíveis para humanos.

Da mesma forma, este trabalho apresenta uma abordagem para sumarização automática semi-extrativa de textos que use diferentes abordagens complementares para sumarização coesiva.

A figura 1 exibe o esquema em blocos da solução aqui apresentada. De maneira geral, a solução é composta de funções intrínsecas, que no conjunto, retornam o resultado esperado, tais como compressão de sumários, fluência, e avaliação de sumários independente de um resumo de referência por meio de um conjunto de medidas do estado da arte (DALE; CHALL, 1948; KINCAID et al., 1975; GUNNING, 1952; COLEMAN; LIAU, 1975; MCLAUGHLIN, 1969; SPACHE, 1953; STEINBERGER; JEZEK, 2009; LOUIS; NENKOVA, 2013a; MCNAMARA et al., 2014; DOWELL; GRAESSER; CAI, 2015a).

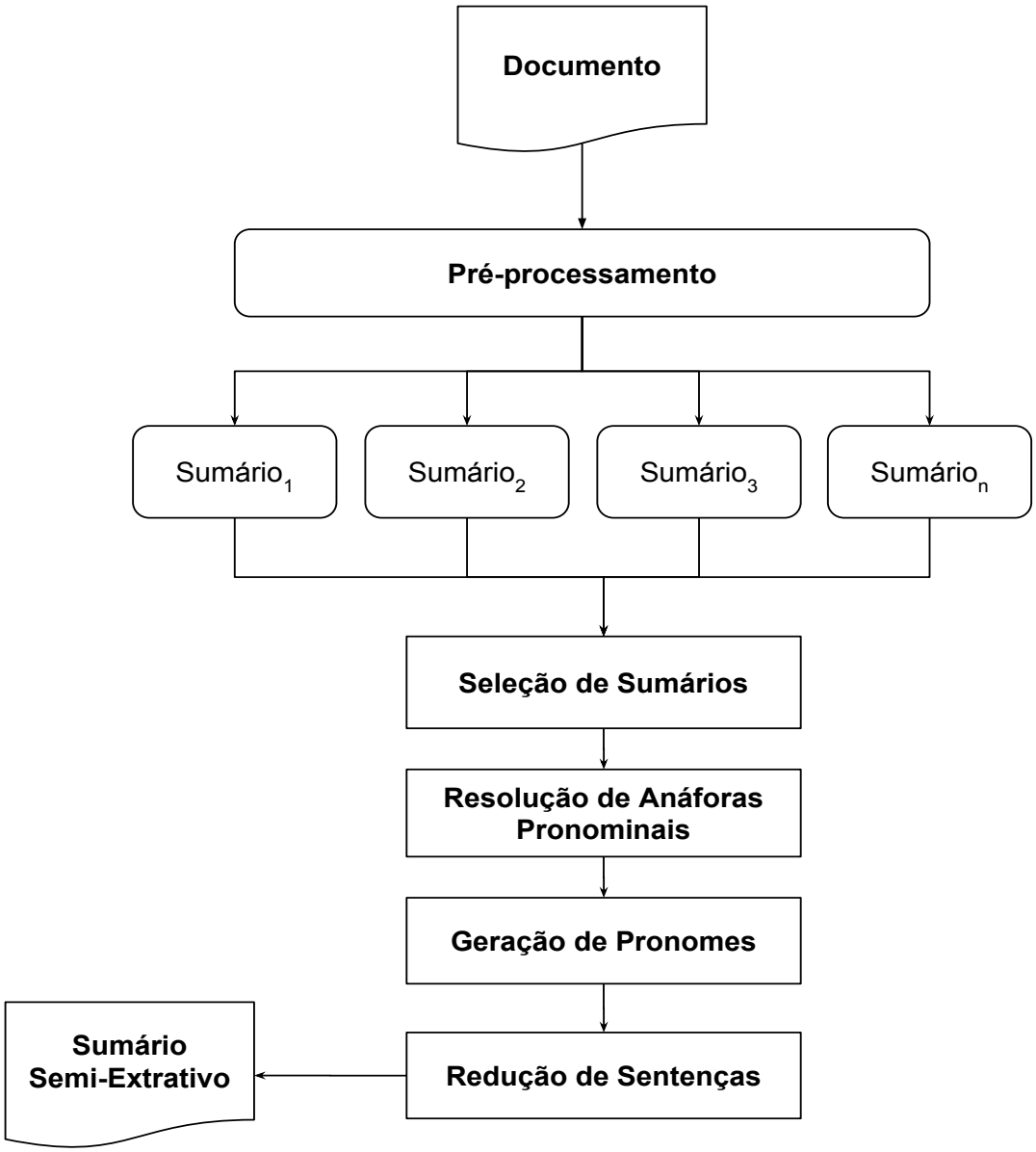


Figura 1 – Visão das abordagens propostas

Cada uma das etapas mostradas na figura 1 são discutidas em detalhes nos próximos capítulos desta tese: (i) Seleção de Sumários é abordada no Capítulo 6; Resolução de Anáforas Pronominais no Capítulo 3; no Capítulo 4 é detalhada a Geração de Pronomes; e por fim, no Capítulo 5 é discutida a Redução de Sentenças.

A ordem dos capítulos é seguida baseada na evolução cronológica da abordagem proposta, por isso as avaliações das medidas para estimar a qualidade de sumários foram as últimas contribuições desta tese.

Para a avaliação das soluções aqui apresentadas, são adotadas medidas de avaliação automáticas e manuais. Os experimentos foram realizados em três diferentes *corpora* de sumarização monodocumento extrativa: *CNN Corpus*, *DUC 2001* e *DUC 2002*. A escolha de tais corpora foi devido ao cumprimento de um dos objetivos dessa pesquisa que é lidar com sumarização genérica.

1.4 CONTRIBUIÇÕES DO TRABALHO

Dentre as contribuições alcançadas neste trabalho, podem-se enumerar:

1. Desenvolvimento de uma metodologia para criação do *Corpus CNN*, mapeando os highlights nas sentenças da notícia. Cada um dos artigos de notícias coletados, foi inicialmente processado para escolher textos auto-contidos e explicativos, cujos textos independessem de vídeos, figuras, links externos, etc. Numa segunda fase, os textos escolhidos na etapa anterior foram re-selecionados para buscar os textos onde existiam os sumários abstrativos apresentados pelos autores originais do artigo de notícia, os chamados *highlights*. Foram então, utilizadas as 22 técnicas de sumarização automática consagradas na literatura para a geração de um sumário extrativo de referência, o (*gold standard*) para cada um dos textos dos documentos originais. Os gold-standards foram avaliados por um grupo de especialistas, seguindo rigorosa metodologia onde cada sumário foi contraposto ao texto original por pelo menos três dos oito especialistas encarregados dessa tarefa, que deveriam concordar com a relevância das sentenças selecionadas. Textos onde não houve o consenso dos especialistas sobre as sentenças escolhidas para o sumário extrativo foram descartados. Ao final desta etapa resultaram os 3.000 documentos finais do Corpus CNN; A escolha do Corpus CNN como base dos experimentos aqui descritos fica assim plenamente justificada uma vez que possivelmente esse seja o maior e melhor corpus até a presente data desenvolvido para a análise de técnicas de sumarização automática.

2. Avaliação quantitativa e qualitativa de sistemas de sumarização automática de texto. Para isso foi necessário automatizar 22 sistemas comerciais e gratuitos, mencionados no item anterior, a fim de avaliar o desempenho de cada um deles em um grande *corpus* de sumarização. Na etapa de experimentos, foram gerados 66 mil sumários para avaliação quantitativa dos sistemas selecionados.

3. Desenvolvimento de uma abordagem independente de sistema de sumarização para resolução de anáforas pronominais, contendo: (i) um componente para identificação de expressões anafóricas em sumários extrativos. Tal componente possibilitou explorar um dos principais problemas de coesão em sumários extrativos. Além disso, esse componente poderá ser usado para avaliar o nível de qualidade de um sumário a partir do número de pronomes não conectados, ou seja, sem uma entidade (por exemplo, um nome) a que ele se refere; (ii) um componente para resolução de anáforas no texto fonte (pré-processamento), gerando a partir dele um texto intermediário. Esse texto, com as correferências resolvidas, poderá ser usado por qualquer sumarizador extrativo, assim as sentenças selecionadas serão autocontidas (todos os pronomes conectados). (iii) um componente para resolução de anáforas no sumário extrativo. Esse componente foi inserido na etapa de pós-processamento dos sumários, gerando

assim uma nova versão semi-extrativa com as correferências quebradas corrigidas. Para validação da abordagem proposta, foram gerados automaticamente 189 mil sumários durante a etapa de experimentos. Além disso, aplicaram-se métodos de avaliação automática e manual.

4. Concepção de uma abordagem para reinserção de pronomes em sumários extrativos, gerando assim novos sumários semi-extrativos mais coesos. A solução proposta aplica o método desenvolvido nesta tese para geração de cadeias de correferências pronominais (por exemplo, as conexões entre pronomes e substantivos), em seguida mapeia as entidades repetidas nos resumos e substituem elas por pronomes. Durante a avaliação da abordagem proposta constatou-se que todos os sistemas do estado da arte selecionados geram sumários com entidades repetidas e para validar a proposta usaram-se métodos de avaliação automática e manual.

5. Desenvolvimento de uma abordagem baseada em regras para simplificação de sentenças em tarefa de sumarização, sem perder a informatividade e a legibilidade do resumo gerado. Tais regras foram avaliadas automaticamente e por humanos.

6. Realização de uma avaliação exaustiva de diversas medidas a fim de estimar a informatividade, legibilidade e coesão de um texto. A arquitetura da abordagem proposta é dividida em duas etapas centrais: (i) Geração dos resumos candidatos; e (ii) Seleção do resumo mais informativo e qualitativo. A geração dos resumos candidatos é executada pelos melhores sistemas de sumarização do estado da arte selecionados nesta tese.

1.5 ESTRUTURA DO DOCUMENTO

Esta tese de doutorado está organizada em sete capítulos. No presente capítulo de introdução foi feita uma contextualização do trabalho desenvolvido numa perspectiva histórica e a caracterização da área.

Uma breve descrição dos demais capítulos deste trabalho de doutorado é apresentado a seguir:

- **Capítulo 2: Uma Avaliação Quantitativa e Qualitativa de Sistemas para Sumarização Automática de Texto.** Neste capítulo são apresentados os experimentos realizados para uma avaliação quantitativa e qualitativa de 22 sistemas de sumarização extrativa do estado da arte. Para avaliação qualitativa esses sistemas foram classificados em diferentes dimensões de sumarizacão, já na quantitativa foram usadas medidas clássicas de avaliação de sumários: ROUGE (LIN, 2004) e Intersecção de Sentenças (IS) (MANI, 2001; FERREIRA et al., 2013). Além disso, é apresentado o Corpus CNN com 3.000 mil notícias extraídas do site CNN para sumarização extrativa, desenvolvido durante o desenrolar da presente tese .

- **Capítulo 3: Sumarização Automática Coesa com Resolução de Anáforas Pronominais.** Esse capítulo apresenta a proposta de uma abordagem para correção de anáforas pronominais quebradas, isto é, pronomes sem conexão a um substantivo no sumário. Para isso, faz uso de um método desenvolvido para identificar anáforas pronominais e substituir pela entidade correspondente, melhorando a legibilidade e a coesão do texto. A proposta foi aplicada em dois cenários: pré-processamento e pós-processamento. No primeiro cenário as anáforas pronominais foram resolvidas no texto original, já no segundo foi aplicado na etapa de pós-processamento do sumário, gerando assim um novo sumário semi-extrativo. Além disso, a abordagem proposta foi avaliada exaustivamente de forma automática e por humanos.

- **Capítulo 4: Reinserção de Pronomes em Sumários.** Nesse capítulo é apresentada uma estratégia para geração de pronomes em resumos extrativos. O objetivo dessa proposta é reduzir o número de entidades repetidas nos resumos, melhorando assim sua legibilidade. Assim como no capítulo anterior, têm-se como saída um novo sumário semi-extrativo e a estratégia foi avaliada exaustivamente de forma automática e por humanos.

- **Capítulo 5: Redução de Sentenças.** É apresentado um conjunto de heurísticas para redução de sentenças de sumários extrativos. As heurísticas desenvolvidas buscam produzir novas sentenças mais curtas que representem as sentenças originais sem perder a sua informatividade e legibilidade. Além do mais, foram avaliadas exaustivamente de forma automática e por humanos.

- **Capítulo 6: Medidas de Qualidade de Sumários.** São analisadas diversas medidas de avaliação de sumários que independam de um resumo de referência e que consideram características, como informatividade e fluêcia (legibilidade, coesão e coerência), na ponderação do melhor sumário. A solução proposta, inicialmente, recebe como entrada diversos resumos extrativos produzidos pelos melhores sumarizadores apresentados nos capítulos anteriores. Em uma segunda etapa, as medidas de avaliação são aplicadas para estimar a informatividade e qualidade textual presentes nos resumos candidatos gerados, visando identificar o resumo estimado como mais informativo e fluente. Além disso, as medidas selecionadas foram avaliadas exaustivamente de forma automática e por humanos.

- **Capítulo 7: Considerações Finais.** São aqui 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 a serem seguidas.

2 AVALIAÇÃO DE SISTEMAS PARA SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS

Atualmente, as ferramentas de sumarização extrativas são mais utilizadas devido não só à facilidade de acesso, desenvolvimento e tempo de processamento, mas também por serem muito menos complexas que a sumarização abstrativa. Diante do exposto e em decorrência da quantidade e diferença dos estudos realizados na área, sentiu-se a necessidade de avaliar sob a mesma perspectiva, antes de qualquer outra estratégia, ferramentas de sumarização disponibilizadas por pesquisadores na web, assim como outras de cunho comercial. E, a partir daí, mensurar e criar algo que seja relevante. Sendo assim, buscou-se fazer uma avaliação quantitativa e qualitativa com o maior número de ferramentas de sumarização hoje existentes, avaliando-se o estado da arte. Uma versão inicial desse trabalho foi publicada por Lins et al. (2012), onde foram avaliadas apenas 6 (seis) ferramentas de sumarização: TextCompactor (TEXTCOMPACTOR, 2014), FreeSummarizer (FREE..., 2011), Smmry Smmry (2009), WebSummaryzer (WEBSUMMARIZER, 2014), Interllexer (INTELLEXER..., 2011) e o Compendium (LLORET; PALOMAR, 2013).

Nesta tese, para avaliar um maior número de ferramentas, criou-se uma ferramenta para automatização e extração de sumários, gerando assim um repositório com 66 mil sumários de diferentes sumarizadores. Tais sumários foram gerados usando um *corpus* também criado neste contexto de trabalho, o qual visa preencher lacunas como: a falta de um grande *corpus* com qualidade textual, compreensível por qualquer área, que possua sumários de referência e relevante estatisticamente. Em virtude disso, estendeu-se o *corpus* usado por Lins et al. (2012) para 3.000 mil notícias recentes extraídas automaticamente do site CNN[1] (www.cnn.com), conforme descrito no capítulo introdutório desta tese. A vantagem do uso desse novo *corpus* repousa sobre a qualidade textual e em torno dos destaques (*highlights*) oferecidos para cada texto, um sumário abstrativo, contendo 3 ou 4 sentenças, fornecidas pelo próprio autor. A partir desses sumários abstrativos, foram gerados sumários extrativos para cada notícia através do mapeamento de cada frase dos *highlights* em uma ou duas frases do texto original, gerando assim o sumário de referência, o *gold standard*. Esse processo de geração do sumário de referência foi feito pelo autor desta tese e seus colegas de grupo de pesquisa, seguindo uma rigorosa metodologia. A descrição detalhada do desenvolvimento do corpus CNN encontra-se na referência (LINS et al., 2018).

Este capítulo concentra-se na descrição, implementação, experimentos, análise quantitativa e qualitativa de 22 sistemas de sumarização extrativas do estado da arte, utilizando o mesmo ambiente (*hardware*, *software* e *corpus*). Os resultados foram publicados e discutidos no artigo "*A Quantitative and Qualitative Assessment of Automatic Text Summarization*

[1] *Cable News Network*

Systems" (BATISTA et al., 2015).

As medidas de precisão, cobertura e *f-measure* (BAEZA-YATES; RIBEIRO-NETO, 1999a) fornecidas pelo ROUGE (LIN, 2004) e a medida do número de sentenças do sumário que estão no sumário de referência (IS) (MANI, 2001; FERREIRA et al., 2013) foram utilizadas para realizar a avaliação quantitativa dos sistemas estudados. Já para avaliação qualitativa, os sistemas foram classificados em diferentes dimensões tradicionalmente utilizadas para categorizá-los.

Nas subseções a seguir serão descritos os sistemas avaliados.

2.1 SISTEMAS AVALIADOS

Nos últimos anos houve o aparecimento de diversas plataformas para sumarização automática de textos, sejam extrativos, abstrativos, mono ou multi-documentos. Esta seção apresenta, em ordem alfabética, uma breve descrição de cada um dos vinte e dois sistemas de sumarização aqui avaliados. Os critérios de escolha desses sistemas foram: disponibilidade de acesso ao código, *Application program interface* (API) para consumo do serviço ou disponibilidade de acesso à aplicação (de uso comercial ou não comercial).

Almus: *Automatic Text Summarizer*

O *Almus* (STEINBERGER; JEŽEK, 2005) afirma produzir sumários extrativos genéricos em inglês, mono ou multidocumento. Ele baseia-se na análise de semântica latente (*latent semantic analysis*) (GONG; LIU, 2001) para seleção das sentenças candidatas ao sumário. Os autores do Almus apresentaram duas desvantagens da abordagem seguida:

- O baixo número de tópicos significativos selecionados;
- Sentenças com alta pontuação, porém se não atingirem um limiar não são selecionadas para o sumário candidato, mesmo que seu conteúdo seja relevante.

Almus lida com esses problemas, propondo uma fórmula baseada no comprimento da sentença e um novo parâmetro independente do número de sentenças do resumo.

Para lidar com sumarização multidocumento (STEINBERGER et al., 2007), o Almus fornece um pacote de serviços básicos. A principal diferença para a sumarização monodocumento é que ele tenta resolver o problema de redundância, eliminando sentenças de alta similaridade no resumo final. A semelhança é medida pela similaridade do cosseno entre os conjuntos de termos das sentenças. O Almus propôs um limiar de 0,6 para eliminar as sentenças de alta similaridade.

Autosummarizer (AutoS)

Autosummarizer (AUTOSUMMARIZER, 2014) é um serviço online para produzir resumos monodocumento extrativos baseado no *ranking* das sentenças mais importantes. Ele usa

um algoritmo para selecionar as sentenças mais importantes do documento original para criar um resumo.

A principal desvantagem do serviço está na impossibilidade de selecionar a taxa de compressão para os resumos. Em geral, um sumário com quatro sentenças é gerado.

Aylien Text Analysis API (Aylien)

A Aylien Text Analysis API (AYLIEN, 2011) é uma ferramenta comercial que fornece diversos serviços de processamento de linguagem natural, incluindo sumarização de textos, classificação de textos, análise de sentimentos, dentre outros.

O Aylien trabalha com sumarização genérica, monodocumento e em inglês. Entretanto, possui um método experimental para alemão, francês, italiano, espanhol e português. Além disso, ele fornece um conjunto de *Software Development Kits* (SDKs) para diferentes linguagens de programação como Java, Python, Ruby, Node.js, PHP e .NET.

Classifier4J (C4J)

Classifier4J (NICK, 2003) é uma biblioteca desenvolvida na linguagem Java, inicialmente criada para fazer a classificação de textos. No entanto, ela também oferece serviços adicionais, como sumarização extrativa monodocumento baseada no método de frequência de palavras (*word frequency*) (FERREIRA et al., 2013). C4J seleciona as sentenças que contêm as palavras mais frequentes no documento para compor o resumo.

Na web há uma extensão da C4J para plataforma .Net, a ferramente é conhecida como *NClassifier* (WHITAKER, 2004).

Compendium

O *Compendium* (LLORET; PALOMAR, 2013) é um sistema para sumarização mono ou multidocumento, capaz de gerar os tipos mais comuns de sumários (baseados em consultas, análise de sentimento, extrativos e abstrativos). As principais funcionalidades dele são:

- Uso de vinculação textual para evitar informações redundantes nos resumos;
- Combinação de estatísticas e técnicas baseadas em cognição para detectar informações relevantes; e
- Geração de sumários abstrativos.

O *Compendium* executa as seguintes etapas:

- Uma análise de superfície linguística (*tokenizer*, *sentence splitter*, *POS-tagging*, *stemming*, identificação de *stopwords*);
- Detecção de redundância (usando para isso a técnica de vinculação textual);

- Identificação de tópicos (identificar os principais tópicos de um documento);
- Detecção de relevância (este estágio atribui um peso para cada frase, dependendo de como ela é relevante dentro do texto, usando *o princípio da quantidade de código*);
- Geração do sumário (as sentenças mais importantes, ou seja, aquelas com pontuações mais altas, são selecionadas e extraídas).

Como o foco neste capítulo é avaliar a qualidade da sumarização extrativa, foi utilizada nos experimentos a versão extrativa do *Compendium*, usando identificação do assunto e detecção de relevância (princípio da quantidade de código e frequência de palavras) para selecionar as sentenças mais importantes para gerar o sumário.

Custom Writing (CW)

Custom Writing (WRITING, 2006) fornece um serviço de sumarização extrativa monodocumento baseado na identificação de palavras-chave (*keywords*). As palavras-chave são selecionadas com base na frequência que aparecem no texto (FERREIRA et al., 2013). O CW funciona apenas para textos em inglês.

Free Summarizer (FS)

O *Free Summarizer* (FREE..., 2011) é um serviço online gratuito que cria sumários monodocumento extrativos, baseados nas frequências de palavras. Ele permite que o usuário selecione a quantidade desejada de sentenças do resumo. Sua desvantagem é que há limite de processamento de arquivos com mais de 15 mil caracteres e a estrutura do documento não é levada em consideração.

HP-UFPE Functional Summarization (HP-FS)

HP-UFPE Functional Summarization (FERREIRA et al., 2013; FERREIRA et al., 2014) é um sistema de sumarização baseado na análise de 17 técnicas encontradas na literatura. Essas técnicas foram amplamente avaliadas em diferentes tipos de documentos, como notícias, blogs e artigos científicos. A melhor combinação de técnicas para cada tipo de documento foi selecionada para compor um sistema híbrido, que primeiramente classifica os documentos por tipo, em seguida, faz-se o resumo usando a combinação que melhor se adéqua ao tipo de documento. Essa etapa trata os problemas de sumarização genérica.

Neste trabalho, HP-FS utiliza a melhor combinação de métodos encontrados em Ferreira et al. (2014) para documentos de notícias. Os métodos utilizados são: *Term Frequency-Inverse Document Frequency* (TF-IDF), *lexical similarity*, *sentence position* e *resemblance to the title*.

- *TF-IDF:* Este algoritmo é dividido nas seguintes etapas: (I) remove todas as *stop words*; (II) calcula a fórmula de TF-IDF apresentada em (FERREIRA et al., 2013) para cada palavra do texto; (III) para cada sentença, soma-se a pontuação do TF-IDF de cada palavra; e (IV) normaliza-se a pontuação.

- Similaridade léxica (*Lexical similarity:*) Baseia-se na suposição de que as sentenças importantes são identificadas por cadeias fortes (*strong chains*).

- Posição da sentença no texto (*Sentence position:*) Este algoritmo classifica as sentenças da seguinte forma: a primeira sentença em um texto tem pontuação de 5/5, a segunda sentença tem pontuação de 4/5, e assim por diante. O mesmo procedimento deve ser aplicado para as últimas sentenças: a última recebe uma pontuação de 5/5, penúltima tem pontuação de 4/5, e assim por diante. Essas etapas são aplicadas para todo o texto.

- Semelhança entre sentença e título (*Sentence resemblance to the title:*) corresponde à sobreposição (*overlap*) de vocabulários entre as sentenças com o título do documento.

Essas técnicas foram combinadas usando a média aritmética dos resultados individuais. Em outras palavras, cada técnica devolve um valor entre 0 e 1, a combinação soma esses valores e divide pelo número de técnicas testadas, que neste caso são quatro. As N sentenças com as maiores pontuações são selecionados para compor o sumário candidato, onde N depende da taxa de compressão definida.

A plataforma HP-UFPE Functional Summarization também tem duas extensões: a primeira gera sumários independentes de idioma (CABRAL et al., 2014), e a segunda trabalha com sumarização multidocumento (FERREIRA et al., 2014).

OpenText Summarizer (OTS)

OpenText Summarizer (OTS) (ROTEM, 2003) é um sistema de sumarização que combina técnicas de PLN com métodos estatísticos convencionais baseado na frequência de palavras para gerar sumários extrativos. OTS incorpora técnicas de PLN através de cadeias léxicas no idioma inglês com sinônimos de termos, bem como regras para *stemming* e *parsing*. Essas técnicas são combinadas com o método estatístico baseado na frequência de palavras para pontuar uma sentença. OTS suporta textos em outros idiomas além do inglês: alemão, espanhol, russo, hebraico, esperanto, dentre outros.

Py Teaser (PT)

Py Teaser (TEASER, 2013) é um sistema de sumarização monodocumento que combina processamento de linguagem natural e aprendizagem de máquina para produzir resumos extrativos. Ele usa recursos como semelhança com o título, tamanho da sentença, posição da

sentença no texto e palavras-chave para selecionar as sentenças mais relevantes (FERREIRA et al., 2013). Originalmente, o PT foi implementado na linguagem *Scala* [2], mas atualmente há uma versão em *Python*. Há suporte para os idiomas inglês, espanhol e russo.

Sumy

Sumy (SUMY, 2015) é um sistema de sumarização monodocumento extrativa multi-idiomas: tcheco, inglês, francês, alemão e eslovaco. Gera resumos usando a frequência das palavras, o algoritmo de *TextRank* (MIHALCEA; TARAU, 2004) e a análise semântica latente (*latent semantic analysis*) (GONG; LIU, 2001). Sumy seleciona as sentenças para compor o resumo usando um desses métodos

SweSum: *Automatic Text Summarizer*

SweSum (DALIANIS; AL., 2003; HASSEL; DALIANIS, 2003) é um ferramenta de sumarização extrativa originalmente desenvolvida para o idioma sueco, mas foi estendida para dinamarquês, espanhol, francês, inglês, alemão, italiano, grego e persa. O método aplicado combina análise estatística e linguística, medindo a importância de uma sentença como a seguir:

Resemblance to the title: Sentenças que contêm palavras do título têm maior pontuação.

Word Frequency: Termos mais frequentes no documento são mais importantes.

Sentence Position: Sentenças no início ou no final do texto são mais relevantes.

Sentence length: O comprimento das sentenças indica quais são as mais importantes.

Average lexical connectivity: O número de termos compartilhados com outras sentenças, presupondo que uma sentença que compartilha mais termos com outras sentenças seja mais importante.

Numerical data: Sentenças que contêm dados numéricos recebem alta pontuação.

Todos os métodos apresentados são combinados usando uma função de ponderação flexível. O sistema gera sumários monodocumento ou multidocumento.

Text Compactor (TC)

Text Compactor (COMPACTOR, 2015) é uma ferramenta de sumarização online e gratuita, criada por Keith Edyburn para Knowledge by Design, Inc. Para gerar um resumo, o TC calcula a frequência de cada palavra no texto. Então, a pontuação é calculada com base

[2] Essa implementação é chamada de *Text Teaser*

no número de ocorrência das palavras mais frequentes para cada sentença. As sentenças mais importantes são as que possuem o maior número de palavras com alta frequência.

Text Compactor diz trabalhar com textos expositivos, como livros didáticos e material de referência, e não é recomendado para ficção (por exemplo, histórias sobre pessoas imaginárias, lugares e eventos).

A ferramenta funciona *online*: o usuário envia um arquivo no formato *.txt* e a saída é um sumário extrativo gerado no mesmo formato. As principais desvantagens desse sistema são: não suporta arquivos longos de entrada (superiores a 15.000 caracteres) e a estrutura do documento não é levada em conta.

Tools4Noobs (T4N)

Tools4Noobs (TOOLS4NOOBS, 2015) é um sumarizador monodocumento extrativo, usa a combinação de técnicas de processamento de linguagem natural com métodos estatísticos convencionais baseados na frequência de palavras para selecionar sentenças que irão compor os sumários candidatos. Esta ferramenta possui 3 etapas:

(i) Extrair sentenças de um texto de entrada;

(ii) Identificar as palavras-chave no texto e contar a relevância de cada palavra; e

(iii) Identificar as sentenças que possuem as palavras-chave mais relevantes.

As palavras-chave são calculadas com base na frequência de palavra no texto. O serviço fornece sumarização genérica para sumários nos idiomas: inglês, russo, persa, árabe e chinês.

Outros Sistemas Avaliados

Alguns sistemas de sumarização encontrados não explicam seus métodos ou características empregadas na sumarização, porém também foram incluídos na avaliação quantitativa aqui efetuada. Esses sistemas são:

- *Sumplify* (SUMP) (SUMPLIFY, 2015)
- *Findwise Summarizers* (FINDWISE, 2015) - a versão demo online suporta quatro técnicas de sumarização: *Biclique*, *Multiple Kernel Learning*, *Sub Modular Optimization* e *Text Rank*.
- *Text Analysis Online* (TAO) (TAO, 2015)
- *TextSummarization* (TextS) (TEXTSUMMARIZATION, 2015)
- *Baseline first N sentences* (OUYANG et al., 2010): Posição da sentença é uma característica importante para sumarização de texto. A hipótese é que as primeiras

sentenças de um documento são as mais importantes (OUYANG et al., 2010). Essa simples heurística foi usada como *baseline* para comparar os sistemas de sumarização.

2.2 ANÁLISE QUALITATIVA

Os sistemas avaliados foram classificados nas seguintes dimensões:

- **Entrada:** O sistema é capaz de gerar sumários a partir de um ou mais documentos de entrada - monodocumento (*Mono*) ou multidocumento (*Multi*).
- **Saída:** O tipo de sumário gerado é extrativo (*Ext*) ou abstrativo (*Abs*).
- **Propósito:** O sistema produz sumários genéricos ou orientados.
- **Linguagem:** O sistema é monolíngue ou trabalha com múltiplos idiomas.
- **Licença:** a ferramenta é comercial ou gratuita.
- **Plataforma:** A plataforma é *desktop* ou está disponível na *Web*.

Para facilitar a apresentação das avaliação nas próximas tabelas deste capítulo, na tabela 3 estão listadas as abreviações para os nomes dos sistemas de sumarização avaliados.

A tabela 4 mostra à classificação dos sistemas usando as dimensões apresentadas.

Analisando a tabela 4 conclui-se que:

1. A maioria dos sistemas são monodocumento. Isso indica que há muitos problemas a serem tratados na sumarização de multidocumento.
2. Somente o sistema Compendium consegue gerar um resumo abstrativo. Propõe, por exemplo, a aplicação de técnicas de fusão de sentenças para transformar um resumo extrativo em abstrativo.
3. Em geral, os sistemas criam resumos genéricos. Em outras palavras, os sistemas analisados criam resumos contendo as informações essenciais do texto sem informações adicionais. As exceções são os sistemas Almus e Compendium.
4. Seis sistemas tentam resolver o problema de sumarização multi-idioma. Conforme apresentado na descrição dos sistemas (Seção 2.1), em geral, cada sumarizador é especializado em idiomas específicos.
5. O único sistema comercial avaliado foi o de Aylien, pois outros resumos comerciais não possuem versões de testes gratuítas.
6. Existem sistemas criados para plataforma desktop e web.

Tabela 3 – Abreviações utilizadas para os sistemas testados.

Systema	Abreviação
Almus	Almus
Auto Summarizer	AutoS
Aylien	Aylien
Baseline First N Sentences	FirstNSent
HP-Functional Summarization	HP-FS
Classifier4J	C4J
Compendium	Compend
CustomWriting	CW
Findwise Biclique	FindBI
Findwise MultipleKernelLearning	FindMKL
Findwise Submodular Optimization	FindSMO
Findwise TextRank	FindTR
Freesummarizer	FS
Open Text Summarizer	OTS
Pyteaser	PyT
Sumplify	Sump
Sumy	Sumy
SweSum	SweSum
TAO Simple Text	TAO
Text Compactor	TC
Text Summarization	TextS
Tools4Noobs	T4N

Tabela 4 – Classificações dos sistemas avaliados com base em diferentes dimensões de sumarização.

Sistema	**Entrada**	**Saída**	**Propósito**	**Idiomas**	**Licença**	**Platforma**
Almus	Mono/Multi	Ext	Genérica/atualização	Inglês	Gratuito	Desktop
AutoS	Mono	Ext	Genérica	Inglês	Gratuito	Web
Aylien	Mono	Ext	Genérica	Multi	Comercial	Web
Classifier4J	Mono	Ext	Genérica	Inglês	Gratuito	Desktop
Compendium	Mono/Multi	Ext/Abs	Genérica	Inglês	Gratuito	Web
			Baseado consulta			
			Análise sentimento			
Custom Writing	Mono	Ext	Genérica	Inglês	Gratuito	Web
HP-FS	Mono/Multi	Ext	Genérica	Multi	Comercial	Desktop
Free Summarizer	Mono	Ext	Genérica	Inglês	Gratuito	Web
OText Sumzer	Mono	Ext	Genérica	Multi	Gratuito	Desktop/Web
Py Teaser	Mono	Ext	Genérica	Multi	Gratuito	Desktop
Sumy	Mono	Ext	Genérica	Multi	Gratuito	Desktop
SweSum	Mono/Multi	Ext	Genérica	Multi	Gratuito	Web
Text Compactor	Mono	Ext	Genérica	Inglês	Gratuito	Web
Sumplify	-	Ext	Genérica	-	Gratuito	Web
Tools4Noobs	-	Ext	Genérica	Multi	Gratuito	Web
Findwise Sumzers	Mono/Multi	Ext	Genérica	-	Gratuito	Web
Text Analysis Online	Mono	Ext	Genérica	-	Gratuito	Web
TextSummarization	-	Ext	Genérica	-	Gratuito	Web

2.3 AVALIAÇÃO DE DESEMPENHO DOS SISTEMAS DE SUMARIZAÇÃO

Esta seção descreve:

(i) o *corpus* utilizado;

(ii) a metodologia seguida nos experimentos para avaliar os resumos;

(iii) os resultados da avaliação de desempenho.

2.3.1 Corpus de Avaliação: Corpus CNN

Todos os experimentos relatados neste capítulo são do domínio de artigos de notícias, utilizando o corpus CNN. A primeira versão desse corpus foi descrita na referência (LINS et al., 2012). Como já dito, a versão atual engloba 3.000 artigos de notícias escritos em inglês coletados no site da CNN[3] contendo notícias de todo o mundo, descritas utilizando linguagem de alto-nível vocabular e gramatical. Todos os textos selecionados para o corpus CNN são auto-contidos, não fazendo referências a tabelas, figuras, ou *links* externos, tendo sido necessário varrer uma quantidade de textos quase dez vezes maior que os selecionados para que fossem escolhidos artigos com tal característica. Os documentos abordam temas gerais originalmente classificados pela CNN como: Cotidiano, Entretenimento, Esportes, Estados Unidos, Justiça, Mundo, Negócios, Opinião, Política, Saúde, Tecnologia e Viagens. A tabela 5 apresenta a distribuição dos textos segundo tal categorização. Como também já aqui descrito, cada artigo tem seus *highlights* correspondente a um resumo abstrativo de alta qualidade, conciso, composto por até quatro sentenças escritas pelos próprios autores das notícias.

Os *highlights* foram usados para guiar um processo semi-automático para criar um sumário de referência extrativo (*gold standard*) para cada documento do corpus CNN. Esse processo foi realizado por anotadores humanos que mapearam cada sentença dos *highlights* em uma ou mais sentenças do artigo original. Um grupo de seis anotadores humanos proficientes na língua inglesa, mas não nativos, foi designado para executar essa tarefa de mapeamento. Para garantir a qualidade dos resumos gerados, dois anotadores mapearam cada documento e, no caso de divergência, um terceiro anotador conduziu o processo de resolução de divergência. O conjunto de sentenças mapeadas pelos anotadores constitui o sumário de referência (*gold standard*), esse sumário pode ser usado para avaliar os sumários extrativos gerados pelos sistemas de sumarização. Os sumários de referência do corpus CNN contêm 10.755 sentenças, o que representa aproximadamente a 10% do total de sentenças dos documentos. A tabela 6 mostra um exemplo de um *highlights*, sumário extrativo e sua versão semi-extrativa de um artigo de notícia da CNN.

Como as sentenças mapeadas foram extraídas sem qualquer modificação, problemas relacionados à coesão dos resumos extrativos ainda podem ser encontrados. Buscando

[3] http://www.cnn.com

Tabela 5 – Categorias de Notícias do Corpus CNN.

Categoria	#Documentos
Cotidiano	98
Entretenimento	241
Esportes	148
Estados Unidos	160
Justiça	224
Mundo	988
Negócios	161
Opinião	192
Política	195
Saúde	290
Tecnologia	132
Viagens	171
Total	3.000

aliviar esse problema, uma versão semi-extrativa do sumário de referência foi criada, realizando o processo de resolução de correferência, seguido de validação humana.

Na avaliação quantitativa descrita na Seção 2.3.3, os sumários de referência semi-extrativos foram utilizados para as avaliações usando a medida ROUGE e os extrativos para a medida do número de casamentos de sentenças entre os sumários extrativos e os sumários de referência (*gold standards*). A tabela 7 fornece estatísticas básicas do corpus CNN.

Estatísticas básicas do Corpus CNN. A tabela 7 resume alguns dados estatísticos do corpus CNN, *highlights* e sumários de referências.

2.3.2 Metodologia de Avaliação

Todos os experimentos para avaliação quantitativa foram realizados no contexto das tarefas de sumarização genérica monodocumento de artigos de notícias escritos em inglês. Na sumarização monodocumento foi usado o corpus CNN, descrito na seção anterior.

- **Recall-Oriented Understudy for Gisting Evaluation** (ROUGE) (LIN, 2004). O processo de avaliação aqui executado adotou a medida ROUGE-2, porque, segundo Lin (2004) possui uma alta correlação com avaliações realizadas manualmente por avaliadores humanos. Como sugerido por Hong et al. (2014), também foram computadas as medidas de cobertura do ROUGE-1 e ROUGE-4, já que elas fornecem maior cobertura e precisão, respectiva, para indentificar resumos informativos. Sendo assim, os experimentos foram conduzidos as medidas de cobertura, precisão e *f-measure* do ROUGE-2 e cobertura do ROUGE-1 e ROUGE-4, usando o algoritmo de stemming e

Tabela 6 – Exemplo de um *highlight*, sumário extrativo e semi-extrativo de referência de um artigo de notícia do corpus CNN.

Modelos	**Sumários**
Highlights	1. Coca-Cola – the world's ubiquitous brown fizzy drink – is staying afloat as the soda market shrinks. 2. Many point to a marketing strategy around the so-called "secret recipe" as key to its resilience. 3. It's never been patented, to keep the formula secret, but many say they have discovered the recipe.
Extrativo de Referência *Gold Standard*	1. Coca-Cola – the world's ubiquitous brown fizzy drink – is staying afloat as the soda market shrinks, and many point to a marketing strategy around the so-called "secret recipe" as key to its resilience in a struggling industry. 2. "They kept the formulas secret, partly in order to increase sales with a sense of special mystery and to prevent competition, but also to keep people from knowing how cheap the ingredients were and how large the profits", **he** says. 3. The company has never patented the formula, saying to do so would require its disclosure.
Semi-Extrativo de Referência	1. Coca-Cola – the world's ubiquitous brown fizzy drink – is staying afloat as the soda market shrinks, and many point to a marketing strategy around the so-called "secret recipe" as key to its resilience in a struggling industry. 2. "They kept the formulas secret, partly in order to increase sales with a sense of special mystery and to prevent competition, but also to keep people from knowing how cheap the ingredients were and how large the profits", **pharmacist John Pemberton** says. 3. The company has never patented the formula, saying to do so would require its disclosure.

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: $-n2-m-fA$, 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): Esta medida computa a intersecção de sentenças entre os resumos gerados automaticamente e

Tabela 7 – Estatísticas Básicas do Corpus CNN.

Textos	
Número de textos de notícias	3.000
Número de sentenças	115.396
Média de sentenças por artigo	38,47
Número de *tokens*	2.296.693
Média do número de tokens por sentença	19,90
Highlights	
Número de sentenças	10.674
Média de sentenças por artigo	3,56
Número de *tokens* nos *highlights*	130.844
Média do número de *tokens* por *highlights*	12,26
Sumário de Referência	
Número de sentenças	10.755
Média do número de sentenças	3,59
Número de *tokens*	269.434
Média do número de *tokens*/sentença	25,05

o conjunto de resumos de referência. Um importante aspecto desta medida é a sua habilidade de identificar métodos com uma alta precisão em reconhecer boas sentenças para compor os resumos. Contudo, a IS só pode ser computada quando resumos extrativos estão disponíveis. Por isso, ela foi executada particularmente no corpus CNN. Na Equação 2.1 é apresenta como a medida IS é computada.

$$IS\left(r_i\right) \frac{S_{resumo}}{S_{referencias}} \tag{2.1}$$

no qual,

- S_{resumo} é o total de sentencas 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 selecionado, caso contrário, o teste de *Wilcoxon signed-rank* (GIBBONS; CHAKRABORTI, 2003) é selecionado; e, por fim, (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 ($M_1 \geq M_2$). Esse 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[4].

2.3.3 Avaliação Quantitativa

A tabela 8 mostra os resultados da medida de interseção de sentenças, a qual corresponde ao número de sentenças selecionadas automaticamente pelos sistemas para criar os resumos extrativos que estão presentes nos sumários de referência criados por humanos. As 3.000 notícias do corpus CNN usadas nesta avaliação têm um total de 10.753 sentenças nos sumários de referência. Os sistemas que atingiram as 5 maiores pontuações foram, respectivamente: AutoS (3.029), C4J (3.014), FirstNSent (2.882), HP-FS (2.842), PyT (2.825) e Aylien (2.657).

Tabela 8 – Número de sentenças nos sumários candidatos que estão presentes nos sumários *gold standard*.

Sistemas	#Sentenças selecionadas corretamente
Human GoldStandard	10.753
AutoS	3.029
C4J	3.014
FirstNSent	2.882
HP-FS	2.842
PyT	2.825
Aylien	2.657
TAO	2.513
TextS	2.481
Almus	2.406
FS	2.399
Sump	2.224
SweSum	2.145
TC	2.137
Sumy	2.090
Compendium	2.064
OTS	2.060
FindSMO	2.052
T4N	1.994
CW	1.901
FindBI	1.851
FindMKL	1.825
FindTR	1.810

[4] https://www.r-project.org/

Os dois melhores sistemas, AutoS e C4J, obtiveram resultados muito semelhantes com apenas 0,50% de diferença. Na verdade, os cinco melhores sistemas obtiveram resultados relativamente próximos, diferindo apenas dentro de um intervalo de 7,72%. Comparando o C4J com FindTR (sistema com pior resultado), foi encontrada uma diferença de 67,35%. Esses resultados mostram que os sistemas avaliados criam uma ampla variedade de resumos.

O sistema usado como *baseline* (*FirstNSent*) mesmo sendo uma simples heurística teve um bom desempenho. Isso corrobora os trabalhos anteriores que apontaram a importância da posição da sentença como característica determinante da sua importância, principalmente em artigos de notícias (OUYANG et al., 2010; FERREIRA et al., 2013).

A acurácia geral foi baixa para todos os sistemas, o AutoS atingiu 28,16%. Isso acontece porque em documentos de notícias os autores usam uma grande redundância, ou seja, várias sentenças apresentam (ou reforçam) o mesmo assunto. Assim, há a possibilidade do sumarizador selecionar sentenças ligeiramente diferentes das escolhidas nos sumários de referência; no entanto, semelhantes no significado.

Como dito na seção anterior, o ROUGE-2 tem a maior concordância com as avaliações manuais, sendo assim, os resultados detalhados de precisão, cobertura e *f-measure* (ordenados por *f-measure*) são mostrados na tabela 9.

Para cada medida, um sistema diferente alcançou o melhor resultado. AutoS (39,29%) alcançou a melhor cobertura, e o FirstNSent (36,25%) a melhor precisão, ambos os sistemas obtiveram resultados estatisticamente superiores aos demais. C4J (34,95%) e FirstNSent (33,55%) obtiveram os dois melhores resultados de *f-measure*, apesar da diferença númerica entre eles, não foram estatisticamente significantes. Porém, em comparação com os demais sistemas, C4J e o FirstNSent mostraram resultados estatisticamente superiores.

Aylien e HP-FS também alcançaram resultados interessantes, com uma diferença de apenas 2,38% a menos que o sistema C4J para a medida *f-measure*.

Com base nos dois experimentos realizados, concluímos que C4J é a melhor ferramenta de sumarização em nossa avaliação, pois atingiu o melhor desempenho em Rouge-2 *f-measure* (tabela 9) e o segundo melhor desempenho no primeiro experimento (tabela 8) com uma pequena diferença de 0.50% em relação ao AutoS (melhor sistema).

O C4J fornece uma sumarização de monodocumento extrativa com base no método de frequência de palavras (FERREIRA et al., 2013). Calcula a frequência de todas as palavras no documento e seleciona as 100 palavras mais frequentes, em seguida, selecionam-se as primeiras sentenças **N** do documento que contêm pelo menos uma das 100 palavras mais frequentes para compor o resumo. **N** depende da taxa de compressão do resumo requerido. Essa abordagem pode ser vista como uma combinação das estratégias de posição de sentenças e método de frequência de palavras.

Os resultados da comparação de desempenho dos sistemas usando cobertura de ROUGE-1, ROUGE-2 e ROUGE-4, ordenados pela cobertura de ROUGE-2 em ordem decrescente, são mostrados na tabela 10. Como mencionado anteriormente, esses resultados de cobertura

Tabela 9 – Comparação do desempenho dos sistemas usando o ROUGE-2 (cobertura, precisão e *f-measure*) (%). O sistema com melhor desempenho é destacado em negrito e o grupo de sistemas estatisticamente similar a ele ($p - valor > 0.05$), se existir, é indicado usando o símbolo †.

Sistema	*Cobertura*	Precisão	*F-Measure*
C4J	35.74	34.19	**34.95**
FirstNSent	31.22	**36.25**	33.55†
Aylien	33.65	31.94	32.77
HP-FS	34.67	30.72	32.57
AutoS	**39.29**	26.94	31.96
FS	32.60	30.22	31.36
T4N	27.71	34.09	30.57
TextS	27.00	33.55	29.92
TC	28.00	31.89	29.81
OPTS	27.75	31.00	29.26
FindSMO	24.87	32.38	28.13
Sump	28.39	27.46	27.91
SweSum	25.86	30.07	27.81
PyT	31.19	23.55	26.84
CW	25.05	25.87	25.45
Sumy	28.00	23.30	25.43
Almus	26.78	23.77	25.18
FindTR	21.73	28.44	24.64
TAO	30.11	20.82	24.62
FindMKL	20.96	28.83	24.28
FindBI	21.01	28.11	24.05
Compendium	21.96	26.14	23.87

são importantes porque o ROUGE-1 e ROUGE-4 fornecem os melhores resultados de precisão e cobertura, respectivamente, para identificar resumos informativos (HONG et al., 2014).

O AutoS alcançou o melhor desempenho de cobertura para Rouge-1, Rouge-2 e Rouge-4. Nas três avaliações realizadas, o AutoS apresenta uma diferença estatística relevante em relação ao segundo melhor sistema. Infelizmente, detalhes das técnicas utilizadas não são encontrados na literatura. No entanto, em outra avaliação de sistemas de sumarização realizada por Jessica Coccimiglio[5], esse sistema também foi bem avaliado pela autora.

É importante notar que os três sistemas mais bem classificados obtiveram o mesmo padrão de resultados mostrados nas tabelas 8 e 9. Esse resultado aponta na direção de que eles são possivelmente os melhores sistemas de sumarização disponíveis hoje.

[5] http://www.makeuseof.com/tag/author/jessicaco/

2.4 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este capítulo apresentou uma análise qualitativa de 22 sistemas de sumarização extrativa do estado da arte sob seis aspectos:

1. *Entrada*: mono ou multidocumento;
2. *Saída*: extrativa ou abstrativa;
3. *Propósito*: sumarização genérica ou orientada;
4. *Linguagem*: mono ou multi-idioma;
5. *Licença*: comercial ou gratuita;
6. *Plataforma*: *desktop* or *web*.

A análise das ferramentas estudadas mostrou que a maioria desses sistemas são: monodocumento, extrativo, genérico, monolíngue, gratuito e estão disponíveis numa plataforma web. Os sumários geradores apresentaram problemas de legibilidade e coesão (correferências quebradas), o que indica as linhas de trabalho para esta tese. Os problemas e dificuldades para obter sumários coesos também foram abordados em outros trabalhos, tais como (SMITH; HENRIK; ARNE, 2012; KASPERSSON et al., 2012; RENNES; JONSSON, 2014; GAMBHIR; GUPTA, 2016). Tal análise também aponta que a principal direção da pesquisa em sumarização é a criação de solução de sumarização para: (i) sumarização multidocumento; (ii) criação de sumários abstrativos; (iii) sumarização orientada, como: baseado em busca e análise de sentimento; e (iv) sumarização multi-Idioma. Nos próximos capítulos, são apresentados os resultados das pesquisas efetuadas pelo autor visando aumentar a taxa de compressão e fluência de sumários. Vários aspectos são estudados, tais como, a redução de sentenças, a resolução de anáforas e a geração de pronomes. Em seguida, avaliam-se os diferentes métodos a fim de se obter um sumário com maior qualidade textual, independente do sistema de sumarização. A aboragem proposta, os experimentos e resultados para diferentes técnicas e sistemas de sumarização ajudam a dar noção da independência defendida nesta tese.

Tabela 10 – Comparação do desempenho dos sistemas usando a medida de cobertura do ROUGE-1, ROUGE-2 e ROUGE-4 (%).

Sistema	ROUGE-1	ROUGE-2	ROUGE-4
AutoS	**58.35**	**39.29**	**33.86**
C4J	51.71	35.74	31.83
HP-FS	51.90	34.67	30.88
Aylien	50.80	33.65	29.54
FS	50.27	32.60	28.54
FirstNSent	44.68	31.22	28.11
PyT	50.28	31.19	26.53
TAO	51.35	30.11	25.73
Sump	46.12	28.39	24.37
Sumy	48.93	28.00	23.74
TC	44.79	28.00	24.27
OPTS	44.91	27.75	23.46
T4N	42.34	27.71	24.40
TextS	40.75	27.00	23.55
Almus	45.99	26.78	22.39
SweSum	42.09	25.86	22.08
CW	44.11	25.05	20.76
FindSMO	40.73	24.87	21.15
Compendium	38.42	21.96	17.93
FindTR	38.54	21.73	17.96
FindBI	36.61	21.01	17.46
FindMKL	37.66	20.96	17.09

[1] Em negrito os resultados estatisticamente significativos

3 RESOLUÇÃO DE ANÁFORAS PRONOMINAIS

A tarefa de selecionar informações relevantes e concisas na Internet é um desafio cada vez mais difícil devido à crescente quantidade de dados disponíveis. Meios automáticos para ajudar os usuários a peneirar e gerenciar uma quantidade tão grande de dados são, portanto, de importância crescente. Em particular, os sistemas de sumarização automática de texto (SAT) têm sido apontados como ferramentas possíveis de identificar e apresentar informações de maneira clara e concisa. Assim, SAT recebeu considerável atenção por parte dos pesquisadores nos últimos anos. Essencialmente, SAT é o processo de criar mecanicamente uma versão resumida de um ou mais documentos relacionados, extraindo suas informações centrais. De acordo com Lloret e Palomar (2012), as abordagens de SAT podem ser divididas em dois ramos principais: (i) *Sumarização Extrativa* que visa selecionar as frases mais importantes de um conjunto de documentos, copiando-as integralmente para o resumo final; e (ii) *Sumarização Abstrativa*, que tenta resumir documentos parafraseando ou modificando as frases mais bem classificadas, a fim de melhorar a coesão do texto, eliminando redundâncias.

Há um interesse crescente em usar a SAT em várias aplicações de Processamento de Linguagem Natural (PLN), como a indexação e classificação de texto (MANNING; RAGHAVAN; SCHüTZE, 2008). A SAT também pode ser usada como uma etapa preliminar na seleção das principais informações a serem visualizadas em dispositivos móveis, tais como celulares (CABRAL et al., 2015).

As técnicas de sumarização extrativas são as estratégias de SAT mais estudadas na literatura (DAS; MARTINS, 2007; LLORET; PALOMAR, 2012), porém são propensas a criarem resumos fragmentados com frases não auto-contidas, cujo significado depende de seu contexto, ou seja, outras sentenças. Esses sumários fragmentados, geralmente, contêm cadeias de correferências anafóricas quebradas que são conhecidas por introduzir problemas de coesão na sumarização automática de texto (MANI; BLOEDORN; GATES, 1998). As cadeias de correferências são construídas por sistemas de Resolução de Correferência (RC) que são capazes de corresponder todas as referências a uma única entidade em um documento, independentemente de suas possíveis formas sintáticas (NENKOVA; MCKEOWN, 2011). Sistemas de RC geralmente combinam substantivos, frases nominais ou pronomes em um documento. Erros relativos a cadeias de correferências anafóricas quebradas são, no entanto, muito comuns em resumos extrativos, especialmente (não surpreendentemente) em resumos curtos (KASPERSSON et al., 2012); e, em particular, para os sumarizadores que focam na cobertura de conteúdo e desconsideram como as sentenças se relacionam umas com as outras. Sumarizadores extrativos, como apontado por Nenkova (2006), frequentemente pontuam relativamente bem porque são avaliados contra resumos de referência criados por humanos usando medidas que priorizam a cobertura de conteúdo,

como ROUGE (LIN, 2004). No entanto, a coesão de resumo geralmente não é considerada em tais avaliações, pois as medidas de avaliação favorecem a cobertura do conteúdo das informações chaves, desconsiderando o quão bem o texto se encaixa (SMITH; HENRIK; ARNE, 2012). Rennes e Jonsson (2014) realizaram uma investigação de rastreamento ocular da coesão de resumo gerada pelos sistemas SAT. Seu estudo teve como objetivo avaliar como diferentes tipos de erros de coesão afetam a leitura de um texto resumido por um sumarizador automático extrativo. Mais precisamente, usando uma câmera de rastreamento ocular, eles se concentraram na natureza de três tipos diferentes de erros de coesão que ocorrem nos resumos extrativos, a saber: a correferência anafórica errônea, a coesão ou contexto ausente e a correferência anafórica quebrada. A análise estatística dos dados revelou que houve ausência de coesão ou contexto e que a referência anafórica quebrada causou alguma perturbação na leitura dos resumos.

O principal objetivo deste capítulo é abordar esses problemas de coesão na SAT, propondo, implementando e avaliando um método para análise e correção de expressões anafóricas pronominais. O método apresentado pode ser considerado como uma abordagem de sumarização extrativa clássica, mas com vantagem de melhorar a coesão geral dos resumos gerados. Em outras palavras, analisando as cadeias de correferências produzidas pelos *parsers* de linguagem natural, o método aqui apresentado pode tanto identificar como filtrar as cadeias de correferências espúrias, reduzindo assim os erros de coesão que são comuns nas abordagens de sumarização extrativa. Além disso, a solução proposta independe do sistema de sumarização extrativa ou método implementado.

A solução proposta baseada em regras pode ser aplicada para analisar e resolver expressões de correferência quebradas em dois cenários distintos: (a) pós-processamento - melhorar a coesão dos resumos gerados pelos sumarizadores extrativos; (b) pré-processamento - eliminar as correferências das sentenças no texto de entrada antes da sumarização extrativa.

Para avaliar a abordagem proposta, dois experimentos - um quantitativo e outro qualitativo - foram conduzidos em dois cenários de aplicação. O experimento quantitativo tem como objetivo avaliar o impacto do método proposto em relação à questão de seleção de conteúdo (informatividade), adotando as tradicionais medidas: ROUGE (LIN, 2004) e similaridade do cosseno (CS) (DONAWAY; DRUMMEY; MATHER, 2000). Enquanto isso, a avaliação qualitativa realizada manualmente por humanos avaliou a coesão dos resumos gerados após o uso do método proposto. Dezessete técnicas de pontuação e quatro sistemas de sumarização extrativa (selecionados a partir das melhores avaliações descritas no Capítulo 2) foram considerados nos experimentos realizados. Tais avaliações são no domínio de artigos de notícias adotando o corpus CNN (LINS et al., 2018). Este corpus é composto de 3.000 artigos de notícias em inglês, todas com resumos extrativos de referência (*gold standard*), produzidos por três especialistas humanos, conforme descrito na Seção 2.3.1.

As contribuições deste capítulo são:

- Proposta e implementação de um novo método para lidar com expressões anafóricas em tarefas de sumarização extrativa, a fim de melhorar a coesão dos resumos gerados;
- Fornecer uma avaliação comparativa do método proposto em relação a várias técnicas e sistemas de sumarização extrativa.

Os resultados deste capítulo estão publicados no artigo "*Automatic Cohesive Summarization with Pronominal Anaphora Resolution*" (BATISTA et al., 2018).

O restante deste capítulo está organizado da seguinte forma: Seção 3.2 apresenta os passos metodológicos da abordagem proposta visando uma técnica de sumarização mais coesa. A metodologia de avaliação e a descrição do corpus usado estão detalhadas na Seção 3.3. Em particular, várias estratégias e sistemas de sumarização extrativa do estado da arte são avaliados e discutidos em três cenários diferentes na Seção 3.4. As conclusões deste trabalho estão presentes na Seção 3.5.

3.1 TRABALHOS RELACIONADOS

Uma das deficiências dos atuais sistemas de sumarização extrativa é que eles geralmente consideram palavras e sentenças isoladamente, ignorando seu relacionamento. Como resultado, os resumos finais produzidos por tais sistemas tendem a conter frases com referências anafóricas pendentes ou quebradas que dificultam a compreensão do resumo como um todo. Para mitigar este problema, vários sistemas de sumarização automáticos que levam em conta a resolução de referência foram propostos.

Steinberger et al. (2007) propôs dois métodos para explorar a resolução de referência em sumarização automática. A primeira abordagem é baseada na análise de semântica latente (LANDAUER; DUMAIS, 1997), que explora as informações anafóricas extraídas pelo seu sistema de resolução de correferência (GUITAR). A segunda abordagem, se assemelha à proposta neste capítulo, examina resumos procurando expressões anafóricas quebradas. Suas estratégias usam o primeiro elemento da cadeia de correferência como o mais representativo para tratar as expressões anafóricas quebradas. Ambas as abordagens foram avaliadas usando o corpus do DUC 2002[1] e obtiveram desempenho significativamente melhor do que as abordagens que não processavam informações anafóricas.

Gonçalves, Rino e Vieira (2008) introduziram um sistema de sumarização (CorrefSum) que melhora a coesão referencial dos resumos extrativos utilizando o conhecimento sobre as cadeias de correferências. Seu sistema melhora o trabalho de Steinberger ao permitir uma escolha mais flexível da entidade mais representativa em uma cadeia de correferência, em vez de sempre usar a primeira entidade como feito por Steinberger. A avaliação do CorrefSum foi baseada no corpus Summ-it (COLLOVINI et al., 2007) contendo 50 textos de notícias em portugês da Folha de São Paulo.

[1] http://duc.nist.gov/pubs.html#2002.

Gonçalves, Rino e Vieira (2008) e o segunda abordagem proposta por Steinberger et al. (2007) verificam os resumos gerados com o objetivo de tratar as expressões anafóricas quebradas. No entanto, estas abordagens não realizam uma análise preliminar das cadeias de correferências para filtrar as correferências espúrias, como o método proposto neste capítulo faz. De fato, os resultados experimentais preliminares obtidos aqui, usando o sistema de resolução de correferência de Stanford, mostram que quase 50,31% das cadeias de correferências encontradas por esses sistemas não são adequadas para a tarefa de sumarização (Seção 3.4.1).

Orăsan (2009) usou a resolução de anáfora para melhorar um sumarizador baseado na simples técnica de frequência do termo – inverso da frequência nos documentos (TF-IDF). O autor argumenta que as sentenças mais importantes em um texto podem ser determinadas com base na importância das palavras que ele contém. O sumarizador foi avaliado em várias versões do corpus CAST (HASLER; ORăSAN; MITKOV, 2003) geradas por seis sistemas de resolução automática de correferências e por um anotador humano. Os resultados experimentais, avaliados usando a medida de similaridade de cosseno (DONAWAY; DRUMMEY; MATHER, 2000), sugerem que a resolução de correferência pronominal foi benéfica para melhorar a legibilidade dos resumos produzidos. Além disso, quando a versão do corpus gerada por um anotador humano foi usada, o sumarizador produziu os melhores resultados para várias taxas de compressão.

Smith, Henrik e Arne (2012) propuseram o sumarizador COHSUM, que é indiretamente baseado na distribuição de correferências nos textos-fonte. O COHSUM calcula uma pontuação para cada sentença, computando a relação de correferência entre as sentenças. A importância das sentenças foi calculada usando uma variante do PageRank (BRIN; PAGE, 1998). A ideia subjacente de COHSUM é que as sentenças que possuem maior número de correferências são consideradas as mais importantes e, portanto, devem ser selecionadas. Os resumos produzidos pela COHSUM foram avaliados no corpus DUC 2002 usando duas medidas: ROUGE (para cobertura de conteúdo) e coesão (análise do número de expressões anafóricas quebradas) em comparação com o documento de entrada. Os resultados revelaram que o COHSUM apresentou um desempenho comparativamente bom em termos de informatividade e produziu significativamente menos cadeias de correferências quebradas em comparação à outros sumarizadores.

O presente trabalho se difere daqueles propostos por Orăsan (2009), Smith, Henrik e Arne (2012), e a primeira abordagem proposta por Steinberger et al. (2007) no sentido de que todos esses estudos integram resolução de correferência como um fator de ponderação para ranqueamento de sentenças ou como heurísticas adicionais durante o processo de sumarização, já o trabalho aqui proposto aplica a resolução de anáfora no texto-fonte em uma etapa pré-processamento ou de pós-processamento sobre os resumos extrativos, independentemente de um sistema de sumarização.

Christensen, Soderl e Etzioni (2013) propuseram um sistema para produzir resumos

coesos a partir de múltiplos documentos. O sistema proposto, chamado G-FLOW, tenta equilibrar coerência e saliência entre sentenças, estimando o nível de coesão de um sumário candidato. O modelo G-FLOW é essencialmente uma representação (baseada em grafo) das relações discursivas entre sentenças com base em várias pistas de coesão presentes no texto, incluindo frases discursivas, substantivos deverbais e correferências. Os autores usam as menções de correferência como recursos para ponderar (ordenar) as sentenças e conectar os nós do grafo (sentenças). Diferentemente, este trabalho emprega a resolução de correferência para analisar expressões anafóricas e substituí-las pelo referente mais representativo.

Silveira (2015) investigou o impacto dos procedimentos de pós-processamento nos resumos extrativos visando obter resumos coerentes. Ela combinou várias tarefas que modificam e relacionam as sentenças umas às outras, como a simplificação das sentenças, a criação de parágrafos e a inserção de conectores de discurso, reunindo tudo como uma tentativa de melhorar a qualidade do resumo final. Seu método é aplicável somente na etapa de pós-processamento, enquanto o método proposto aqui também pode ser aplicado na etapa de pré-processamento. Além disso, o trabalho de Silveira não usa resolução de correferência.

O método baseado em regras proposto neste capítulo faz um pré-processamento de corpus de entrada ou pós-processamento do resumo, substituindo as correferências pronominais pela entidade mais representativa da cadeia de correferência. O método proposto é independente de sistemas de sumarização extrativa, enquanto os estudos relacionados estão fortemente ligados a um sistema específico. Além disso, o método aqui proposto introduz critérios específicos para tais substituições, evitando repetições de expressões anafóricas no resumo, enquanto os trabalhos relacionados sempre substituem todas as menções pronominais. Finalmente, todos os estudos anteriores não conduziram uma avaliação tão extensa quanto a relatada aqui, que empregou uma metodologia de avaliação envolvendo várias técnicas e sistemas de sumarização extrativa, adotando um corpus de sumarização muito maior.

3.2 SUMARIZAÇÃO AUTOMÁTICA COESA

Como já foi dito, os resumos gerados pelos sistemas de sumarização extrativos geralmente contêm correferências quebradas (SMITH; HENRIK; ARNE, 2012; NENKOVA; MCKEOWN, 2011). Para resolver esse problema, foi proposta uma arquitetura de software flexível (figura 2) que integra um sistema de resolução de correferência do estado da arte e várias regras de filtragem e substituição de correferências quebradas.

A figura 2 mostra a arquitetura funcional do método de resolução de expressões anafóricas - *Anaphoric Expression Solver* (AES). Método proposto para analisar, filtrar e resolver cadeias de correferências anafóricas em textos. O módulo AES pode ser aplicado em dois contextos distintos, mas relacionados (figuras 2(a) e 2(b)).

Na figura 2(a), um conjunto de documentos é dado como entrada para o componente de pré-processamento de texto que, além das subtarefas tradicionais de processamento de linguagem natural, incluindo tokenização e POS *tagging*, também executa RC. RC ainda é uma tarefa muito desafiadora, pois o desempenho dos sistemas do estado da arte está em torno de 60% em termos de *F1-measure*, de acordo com a tarefa compartilhada do CoNLL-2011 (LEE et al., 2011). De fato, um número significativo de falsos positivos e falsos negativos ainda pode ser encontrado mesmo usando os atuais sistemas de RC do estado da arte. Para atenuar esse problema, o componente AES, apresentado na figura 2(a), visa melhorar a análise de saída da etapa de pré-processamento do texto aplicando um conjunto de regras que:

- filtra as cadeias de correferência mais relevantes a tarefa de sumarização;
- encontra as entidades mais representativas (nomes) e seus referentes correspondentes (pronomes);
- corrige muitos tipos de erros cometidos pelos sistemas de Resolução de Correferência.

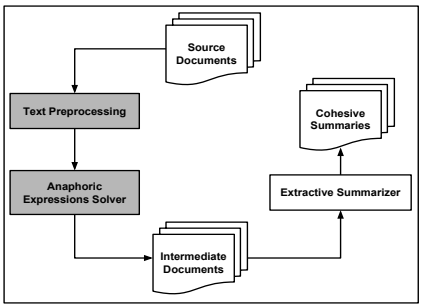
Como resultado, nos documentos originais são inseridas as correferências resolvidas, originando novos documentos, chamados de intermediários nas próximas seções. Estes documentos, contêm informações de correferências mais precisas, podendo então, serem processados por um sistema de sumarização extrativo que produzirá resumos mais coesos.

A figura 2(b) esquematiza a tarefa de combinar os resumos extrativos gerados com os documentos de origem para produzir resumos mais coesos. Estes resumos são gerados após a aplicação das seguintes regras:

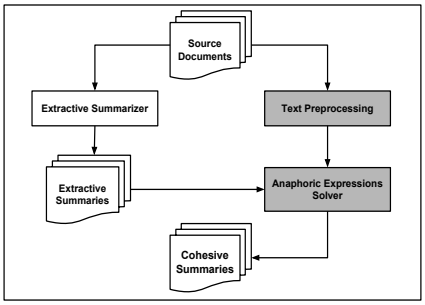
- procura por correferências quebradas no resumo extrativo;
- identifica e extrai as instâncias da entidade cujos referentes estão quebrados no resumo extrativo e encontra a instância mais representativa para tais menções;
- gera uma nova versão do resumo extrativo substituindo os pronomes ou expressões anafóricas pela entidade mais representativa, controlando o número de repetições dessa entidade no resumo final (semi-extrativo).

A motivação desta pesquisa é investigar em que ponto os atuais sistemas e técnicas de sumarização extrativa podem se beneficiar das heurísticas introduzidas pelo método AES nos cenários de aplicação apresentados acima. Outro objetivo da abordagem proposta é investigar a influência da resolução anafórica no desempenho dos sumarizadores extrativos.

Os componentes da arquitetura funcional do AES (figura 2) são detalhados a seguir.



(a) Resolução de Expressões Anafóricas aplicadas ao corpus de entrada



(b) Resolução de Expressões Anafóricas aplicadas ao sumário

Figura 2 – Arquitetura funcional do Método de Resolução de Expressões Anafóricas empregada em dois cenários distintos.

3.2.1 Pré-processamento de Texto

A etapa de pré-processamento de texto fornece a análise morfossintática dos documentos de entrada. A implementação atual do método AES conta com o *Stanford CoreNLP toolkit*, um sistema de *Natural Language Processing* (NLP) de última geração capaz de executar uma infinidade de subtarefas de linguagem natural, incluindo divisão de frases, tokenização, *POS tagging*, dentre outros[2].

As seguintes subtarefas de PLN do CoreNLP foram escolhidas:

- *Divisão de frases:* delimita as sentenças no texto;
- *Tokenização:* identifica as palavras ou símbolos (*tokens*) nas frases;
- *POS tagging:* fornece categorias de *part-of-speech* dos tokens;
- *Lematização:* remove terminações flexionais, como a forma plural de substantivos, retornando a base ou a forma de dicionário de uma palavra;
- *Reconhecimento de Entidade Nomeada:* identifica e classifica uma palavra ou um grupo de palavras consecutivas em uma sentença em categorias pré-selecionadas, como Pessoa, Organização, Local, entre outras;
- *Resolução de Correferência:* descobre todas as entidades relevantes e seus referentes (nominais e pronominais) em um texto.

Devido à sua importância primordial na solução proposta, o sistema de resolução de correferência de Stanford é descrito a seguir.

[2] Stanford Coreference Resolution System. http://nlp.stanford.edu/software/dcoref.shtml.

O Sistema de Resolução de Correferência de Stanford

O Sistema de Resolução de Correferência de Stanford (SCRS) estende o framework *Multi-Pass Sieve* de Lee et al. (2013) que consiste em uma coleção de modelos determinísticos de resolução de correferência que incorpora informações léxicas, sintáticas, semânticas e discursivas. O sistema propaga informações globais compartilhando atributos (por exemplo, gênero e número) entre as menções no mesmo *cluster*.

O SCRS foi selecionado porque fornece desempenho superior em relação aos sistemas de resolucão de correferência do estado da arte, conforme relatado em Lee et al. (2013), sendo classificado em primeiro lugar na tarefa de resolução de correferência do CoNLL-2011 (PRADHAN et al., 2011), resultando em uma pontuação média de 57,8 na trilha fechada[3] e 58,3 na trilha aberta[4].

Embora o SCRS tenha alcançado bons resultados, ainda há espaço para melhorias. Assim, a presente abordagem tenta melhorar o desempenho da SCRS integrando um conjunto de heurísticas como uma das contribuições do método de resolução de expressões anafóricas, descrito a seguir.

3.2.2 Método de Resolução de Expressões Anafóricas

O método de resolução de expressões anafóricas, mostrado na figura 2, é uma das principais contribuições deste capítulo. Ele usa a saída da etapa de pré-processamento de texto e a saída do sistema RC.

A implementação do AES estende os resultados dos trabalhos anteriores (descrito na Seção 3.1) aplicando um conjunto de heurísticas baseadas em regras para as seguintes tarefas:

- **(Tarefa 1)** Filtrando a saída das cadeias de correferências espúrias antes de identificar as entidades mais representativas (MRE) e os referentes; e
- **(Tarefa 2)** Melhorando a qualidade do texto dos resumos gerados. Em outras palavras, reduzir a redundância das entidades mais representativas nos resumos finais. Tal redundância é geralmente devida à estratégia simples de substituir todos os referentes pela entidade correspondente, como feito por Steinberger et al. (2007).

A heurística baseada em regras na Tarefa 1 está fundamentada na noção de MRE em uma cadeia de correferência, definido como a entidade representada pelo nome completo seguida pela mais curta de suas aposições presentes no texto. Outra forma aceitável de determinar a entidade mais representativa consiste em até cinco *tokens* não separados por vírgulas, por exemplo, entidades com várias palavras.

[3] Somente os dados fornecidos podem ser usados, ou seja, WordNet e dicionário de gênero (*gender gazetteer*)

[4] Qualquer fonte de conhecimento externa pode ser usada. Eles usaram como recursos adicionais *animacy*, gênero, *demonym* e dicionários (*gazetteers*) de países e estados

As heurísticas na Tarefa 2 abordam o problema de evitar a resolução de todas as correferências anafóricas que levariam à informações redundantes e entidades repetitivas no resumo gerado. Para melhorar a coesão dos resumos, a Tarefa 2 verifica a distância, em termos do número de sentenças, entre a entidade e seu pronome de referência. Mais precisamente, se a informação da entidade é encontrada na sentença anterior mais próxima, então não há necessidade de resolver a correferência dada pelo pronome encontrado na sentença seguinte. Uma aplicação direta dessa idéia está relacionada à substituição de expressões anafóricas que têm seus contextos não presentes no resumo gerado.

A seguir, é fornecido um exemplo de entidade representativa, documento intermediário e resumo coeso, conforme definido pelo AES .

(Exemplo 1) Notícia em <http://edition.cnn.com/2013/01/24/business/davos-uk-cameron/>: **Cameron: We must focus on trade, taxes, transparency**

S1: Free trade, transparency and a crackdown on tax cheats will be at the heart of Britain's G8 presidency, [**Prime Minister David Cameron**]$_1$ told the World Economic Forum in Davos on Thursday as [*he*]$_1$ set out [*his*]$_1$ vision for a more competitive Europe.

S2: The speech comes a day after [**Cameron**]$_1$ made headlines by promising the British people a vote on European Union membership if [*he*]$_1$ wins the next general election in 2015.

Uma resolução de correferência ideal encontraria a seguinte cadeia de correferência:

Cadeia 1:

he na sentença S1 (Tarefa 1),
his na sentença S1 (Tarefa 1),
Cameron na sentença S2 (Tarefa 2),
he na sentença S2 (Tarefa 1)

MRE: *Prime Minister David Cameron* in S1 (Tarefa 2)

Exemplo de saída do método (*Replace_Document_MRE*):

S1: Free trade, transparency and a crackdown on tax cheats will be at the heart of Britain's G8 presidency, **Prime Minister David Cameron** told the World Economic Forum in Davos on Thursday as *he* set out *his* vision for a more competitive Europe.

S2: The speech comes a day after **Prime Minister David Cameron** made headlines by promising the British people a vote on European Union membership if *he* wins the next general election in 2015.

Exemplo de saída do método (*Replace_Summary_MRE*):

S1: Free trade, transparency and a crackdown on tax cheats will be at the heart of Britain's G8 presidency, **Prime Minister David Cameron** told the World Economic Forum in Davos on Thursday as *he* set out *his* vision for a more competitive Europe.

S2: The speech comes a day after *Cameron* made headlines by promising the British people a vote on European Union membership if *he* wins the next general election in 2015.

Em todos os exemplos acima, pode-se observar que as heurísticas propostas preferem entidades representativas mais curtas, porém mais informativas. As duas tarefas anteriores

são executadas pelo algoritmo AES, descrito a seguir:

Algoritmo 1: Algoritmo de Resolução de Expressões Anafóricas.

```
1  Function Anaphoric_Expression_Solver(C)
2      foreach document D in corpus C do
3          SUMM = summarizer (D)
4          retrieve all coreference chains CC in D
5          filter out CC containing more than one mention
6          foreach CC_i in CC do
7              MRE = Find_Most_Representative_Entity(CC_i);
8              if MRE ≠ null then
9                  Replace_Document_MRE(MRE, CC_i, D);
10                 Replace_Summary_MRE(MRE, CC_i, SUMM);
11             end
12         end
13     end
14 Function Find_Most_Representative_Entity(cc)
15     tempMention = null;
16     foreach mention M in CC do
17         if POS_Tag(M) in [(JJ)* (NN* or NNP*)] then
18             tempMention = M;
19             if (M contains apposition relation) and (M.length ≤ 10) then
20                 return tempMention;
21             end
22         end
23     end
24     return tempMention;
25 Function Replace_Document_MRE(MRE, CC, D)
26     foreach mention M in CC do
27         retrieve sentence S in D containing M;
28         if (S does not contain MRE) then
29             S = replace M by MRE in sentence
30         end
31     end
32 Function Replace_Summary_MRE(MRE, CC, SUMM)
33     foreach mention M in CC do
34         retrieve sentence S in SUMM containing M;
35         if (S does not contain MRE AND the precedent sentences do not contain MRE) then
36             S = replace M by MRE in sentence;
37         end
38     end
```

A função *Find_Most_Representative_Entity* é responsável por encontrar a entidade mais representativa em uma determinada cadeia de correferência, de acordo com os critérios mencionados anteriormente, enquanto a função *Replace_Document_MRE* substitui as expressões anafóricas pelas entidades mais representativa da cadeia em um documento, mas controlando o número de repetições dos referentes. Para encontrar o MRE, o número máximo de 10 tokens para uma entidade foi definido empiricamente para evitar informações desnecessárias.

A função *Replace_Summary_MRE* resolve as correferências quebradas nos resumos gerados pelos sumarizadores extrativos. Ela verifica se a entidade representativa está ou não contida em uma sentença anterior para evitar a redundância no resumo final.

3.2.3 Análise dos Algoritmos de Substituição Correlatos

O algoritmo pós-processamento de resumos (AES) proposto nesta tese difere do proposto por Steinberger et al. (2007) no sentido de que o método aqui impõe restrições adicionais tanto sobre a definição de uma entidade representativa em uma cadeia de correferência quanto sobre como as substituições são realizadas. Mais precisamente, o trabalho de Steinberger sempre seleciona a primeira expressão nominal na cadeia de correferência como sua entidade representativa. No entanto, é provável que essa abordagem introduza expressões de entidade repetitivas que podem influenciar fortemente o comprimento final das sentenças resultantes (veja-se o Exemplo 1). Portanto, como os resultados experimentais realizados nesta tese revelaram, muitos dos sistemas de sumarização extrativa favorecem a seleção de sentenças mais longas.

Outro problema estreitamente relacionado é que as versões mais longas das entidades representativas invariavelmente alteram a distribuição de frequência de seus termos constituintes, o que pode induzir ao erro os sumarizadores baseados em técnicas de frequência de termos ou palavras. O método de substituição aqui proposto escolhe a melhor entidade representativa para uma dada cadeia de correferência, dando preferência entre todas as candidatas, à mais curta e à mais informativa, usando as regras da Tarefa 1, descritas na Seção 3.2.2. Finalmente, o método de substituição proposto por Steinberger et al. (2007) pode introduzir redundância nos resumos finais, contrariamente à solução proposta que aplica heurísticas para controlar o número de repetições da entidade representativa na mesma sentença ou em outras sentenças no mesmo resumo.

3.3 CONFIGURAÇÕES DOS EXPERIMENTOS

Esta seção descreve as configurações dos experimentos que compreendem o corpus (Seção 3.3.1), as medidas usadas para avaliar a solução proposta para a resolução de expressões anafóricas, os sistemas e técnicas de sumarização extrativa utilizadas (Seção 3.3.2) e, por fim, os cenários de avaliação (Seção 3.3.4). Na Seção 3.3.4.1, um cenário de sumarização de referência (*baseline*) que avalia todos os sistemas e técnicas de sumarização no corpus da CNN é apresentado. As Seções 3.3.4.2 e 3.3.4.3 descrevem os dois cenários de aplicação do método proposto (AES): na etapa de pré-processamento ou na etapa de pós-processamento no processo de sumarização. Além disso, as medidas quantitativas e qualitativas usadas para avaliar o método proposto são apresentadas nas Seções 3.3.5 e 3.4.2.1, respectivamente.

3.3.1 O Corpus CNN

Todos os experimentos realizados utilizaram o corpus CNN (LINS et al., 2018) do domínio de artigos de notícias. A seguir, são fornecidas algumas estatísticas básicas e a distribuição dos pronomes encontrados nesse corpus.

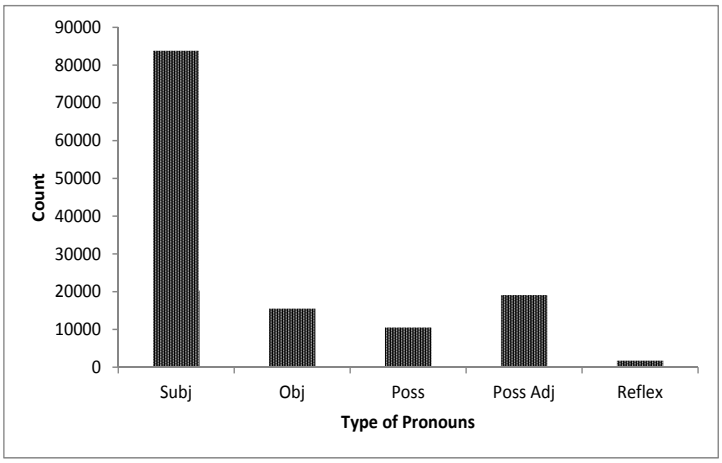


Figura 3 – Frequência de pronomes por tipo encontrados no corpus CNN.

Estatísticas básicas do corpus CNN. Na tabela 11 são apresentadas algumas estatísticas básicas sobre o corpus.

Tabela 11 – Estatística básica do Corpus CNN.

Notícias	
Total de notícias	3.000
Total de sentenças	115.396
Total de *tokens*	2.296.693

Distribuição dos pronomes. A figura 3 apresenta a distribuição dos pronomes no corpus.

A classe morfológica de um dado *token* no corpus CNN é determinada pelo seu *POS tag*. A figura 3 mostra que os pronomes pessoais (*Subj*) (I, you, he, she, it, we, you, and they) são os mais frequentes, seguidos dos pronomes adjetivos possessivos (*Poss Adj*). Os pronomes objeto (*Obj*) são também bastante frequentes. De fato, eles superam em número os possessivos (*Poss*) e os pronomes reflexivos (*Reflex*) juntos. Entre todos os pronomes encontrados no corpus CNN, os mais frequentes são: *it* (18.496), *he* (15.697), *I* (10.487), *his* (10.413), *they* (9.221), e *we* (8.233). O número total de pronomes (130.689), como mostrado na figura 3, corresponde a 5,59% de todos os *tokens* do corpus. Evans (2001) encontrou aproximadamente a mesma porcentagem de pronomes (5,7%) no *British National Corpus* (BNC) (BURNARD, 1995). Tais resultados justificam a escolha do domínio de notícias para avaliar a solução proposta para a resolução de expressões anafóricas, pois os pronomes são encontrados com mais frequência no domínio de notícias do que em outros corpus (BIBER; CONRAD; REPPEN RANDI, 1998; ORăSAN, 2009).

3.3.2 Sistemas e Técnicas de Sumarização Extrativa

Para os três cenários de avaliação da solução proposta para a resolução de expressões anafóricas (Seção 3.2), quatro sistemas de sumarização extrativa foram selecionados:

Autosummarizer (AutoS), Classifier4J (C4J), HP-UFPE Functional Summarization (HP-UFPE) e Aylien Text Analysis API (Aylien). Esses sistemas foram escolhidos devido ao seu bom desempenho, conforme relatado por Batista et al. (2015) e detalhado no Capítulo 2 desta tese. Além disso, as seguintes técnicas foram avaliadas: Aggregate Similarity (AS), Word Co-Occurrence (N-GRAM), Sentence Centrality 2 (SC), Bushy Path (BP), Sentence Length (SL), TextRank Score (TS), Cue-phrase (CP), Sentence Centrality 1 (BLEU), Sentence Position in Paragraph (SPP), Lexical Similarity (LS), Term Frequencies (TF/IDF), Word Frequency (WF), Upper Case (UC), Resemblance to the Title (RT), Inclusion of Numerical Data (ND), Proper Noun (PN) e Sentence Position in Text (SPT). Mais detalhes sobre todas as técnicas de pontuação de sentenças citadas podem ser encontrados na secção 3.3.3.

As técnicas de sumarização acima foram avaliadas com o objetivo de estimar qual produz o maior número de correferências quebradas. Os resultados de tal avaliação esclarecem seu impacto nos dois cenários de avaliação em comparação com o cenário de referência.

Os sistemas de sumarização diferem das técnicas de pontuação de sentença no sentido de que eles podem ser compostos por uma combinação de várias técnicas, ou seja, é uma solução específica com configurações e decisões de design particulares.

A subseção a seguir descreve 17 métodos de sumarização extrativa baseados em pontuação de sentenças amplamente utilizados e referenciados na literatura (FERREIRA et al., 2013), aplicados em sumarização monodocumento ou multidocumento nos últimos 10 anos. Os métodos supracitados são descritos a seguir e foram (re)implementados pelo autor desta tese.

3.3.3 Métodos de Pontuação para Sumarização Extrativa

A primeira referência para sumarização de textos usando pontuação de sentenças remonta a 1958 (LUHN, 1958; LLORET; PALOMAR, 2009). Como já registrado outrora, o foco dessa área de pesquisa é guiado pela seguinte questão: como um sistema pode determinar quais sentenças são representativas do conteúdo de um texto específico? Em geral, três abordagens são utilizadas para pontuação: (i) palavras - atribuição de pontos para as palavras mais importantes; (ii) sentenças - verificando características de sentenças tais como a sua posição no documento, semelhança com o título, entre outras; e (iii) grafos - análise da relação entre sentenças, designando pontuações. Apresentam-se a seguir detalhes sobre os principais métodos utilizados em cada uma das abordagens acima enumeradas.

3.3.3.1 Métodos Baseados em Pontuação de Palavras

Os primeiros métodos em pontuação de sentenças foram baseados em palavras. Cada palavra recebe uma pontuação e o peso de cada frase é a soma de todas as pontuações das suas palavras constituintes. As abordagens contidas na literatura são descritas abaixo.

Frequência de Palavras

Como o nome sugere, as palavras mais frequentes no texto, recebem maiores pontuações (LUHN, 1958; LLORET; PALOMAR, 2009; GUPTA; PENDLURI; VATS, 2011; KULKARNI; PRASAD, 2010; ABUOBIEDA et al., 2012). Em outros termos, sentenças contendo as palavras mais frequentes em um documento terão uma chance maior de serem selecionadas para o sumário. A suposição é de que quanto maior a frequência de uma palavra no texto, mais provável é que ela indique o assunto principal do documento.

Term Frequency—Inverse Document Frequency (TF/IDF)

A hipótese assumida por esta abordagem é que se existem "palavras mais específicas" em uma determinada sentença, então esta sentença é relativamente mais importante. As palavras em questão geralmente são substantivos, exceto para substantivos temporais ou adverbiais (MURDOCK, 2006; SATOSHI et al., 2001). O algoritmo TF/IDF executa uma comparação entre a frequência do termo (tf) num documento (neste caso, cada frase é tratada como um documento) e a frequência do documento (df), o que significa o número de vezes que a palavra ocorre ao longo de todos os documentos. A pontuação TF/IDF é calculada da seguinte forma:

$$\frac{TF}{IDF(w)} = DN\left(\frac{log(1+tf)}{log(df)}\right) \tag{3.1}$$

onde DN é o número de documentos.

Maiúsculas

Este método atribui uma pontuação maior para palavras que contenham uma ou mais letras maiúsculas, segundo Prasad et al. (2012). Pode ser um nome próprio, iniciais, palavras em destaque, entre outros. A pontuação é calculada da seguinte forma:

$$CPTW\,(j) = \frac{NCW\,(j)}{NTW\,(j)} \tag{3.2}$$

onde, $CPTW$ = Razão do total de palavras com primeira letra maiúscula presentes na sentença pelo total de palavras presentes na sentença, NCW = Número de palavras com primeira letra maiúscula, e NTW = Número total de palavras presentes na sentença.

$$UCF\,(j) = \frac{CPTW\,(j)}{Max\,(CPTW\,(j))} \tag{3.3}$$

onde, UCf = é o coeficiente de pontuação utilizado neste método.

Nomes Próprios

Supõe-se que as sentenças que contêm um maior número de nomes próprios têm maior importância no texto; assim, elas são elegíveis a serem incluídas no sumário do documento (FATTAH; REN, 2009). Esta é uma especialização do método Maiúsculas, explicado acima.

Coocorrência de Palavras

Neste caso, a coocorrência de palavras mede a chance de dois termos de um texto aparecerem ao lado de outro em uma determinada ordem. Uma maneira de implementar esta medida é usando n-gramas (MARINO et al., 2006), que é uma sequência contígua de n itens de uma determinada sequência de texto. Em suma, ele dá maior pontuação para as sentenças que possuem maior frequência de coocorrências de palavras (LIU; WEBSTER; KIT, 2009; GUPTA; PENDLURI; VATS, 2011; TONELLI; PIANTA, 2011).

Similaridade Léxica

Esta técnica baseia-se na suposição de que as sentenças importantes são identificadas pela ocorrência de palavras de mesmo significado (sinônimos) ou outra relação semântica, de importância reconhecida (LIU; WEBSTER; KIT, 2009; MURDOCK, 2006), e.g. palavras mais frequentes ou nomes próprios.

3.3.3.2 Métodos Baseados em Pontuação de Sentenças

Esta abordagem analisa as características da própria sentença e foi utilizada pela primeira vez em 1968 (EDMUNDSON, 1969) analisando a presença de palavras usadas como pontos de importância (*cue-phrases*) em sentenças. As principais técnicas que seguem esta linha estão descritas a seguir.

Ponto de Importância na Sentença (*Cue-phrases*)

Em geral, as sentenças que começam por "em suma", "conclui-se", "nossa pesquisa", "o documento descreve", além de frases fortes ou de efeito, e.g. "o melhor", "o mais importante", "de acordo com os estudos/resultados", "significativamente", "importante", "em particular", "dificilmente", "impossível", bem como termos de domínio específico em frases podem ser bons indicadores de quão significante é a sentença para um documento (GUPTA; PENDLURI; VATS, 2011; KULKARNI; PRASAD, 2010; PRASAD et al., 2012). A maior pontuação é atribuída para sentenças que contenham palavras ou frases com pontos de importância, utilizando a fórmula:

$$CP = \frac{CPS}{CPD} \tag{3.4}$$

onde, $CP =$ é o coeficiente de pontos de importância (*cue-phrases*), $CPS =$ são os pontos de importância na sentença, $CPD =$ o total de pontos de importância no documento.

Dado Numérico na Sentença

Normalmente, a sentença que contém dados numéricos é importante e tem alta probabilidade de ser incluída no sumário do documento, segundo as referências (KULKARNI;

PRASAD, 2010; FATTAH; REN, 2009; ABUOBIEDA et al., 2012; PRASAD et al., 2012). Esse tipo de frase normalmente se refere a algumas informações importantes, como data do evento, transação de dinheiro, porcentagem de ganho ou perda, entre outros.

Tamanho da Sentença

Este recurso é utilizado para penalizar sentenças muito curtas (FATTAH; REN, 2009) ou muito longas (ABUOBIEDA et al., 2012), essas frases não são consideradas como uma seleção ideal para o sumário. Para calcular o tamanho da sentença o método usa o número de palavras contidas na frase. Além disso, (SATOSHI et al., 2001) penaliza sentenças que são mais curtas do que um determinado comprimento, ou seja, um limiar é definido. O primeiro caso pode ser calculado deste modo:

$$Pontuacao\,(s) = Tamanho\,(Sentenca) * TamanhoMedio\,(Sentencas) \tag{3.5}$$

Já a penalidade do segundo caso pode ser obtida usando a condição:

$$Penalidade\,(S_i) = \begin{cases} L_i & se\,(L_i > C) \\ L_i - C & senao \end{cases} \tag{3.6}$$

onde, Li = tamanho da sentença i e C = tamanho definido como limiar pelo usuário.

Posição da Sentença

Existem muitas técnicas que usam a posição de sentença como um critério de pontuação (FATTAH; REN, 2009; SATOSHI et al., 2001; BARRERA; VERMA, 2012; ABUOBIEDA et al., 2012; GUPTA; PENDLURI; VATS, 2011). No trabalho de Abuobieda et al. (2012), a primeira sentença do parágrafo é considerada importante e uma forte candidata a ser incluída no resumo. Gupta, Pendluri e Vats (2011) afirmam que as primeiras frases dos parágrafos e palavras nos títulos e subtítulos são relevantes para a sumarização. O método proposto por Satoshi et al. (2001) atribui pontuação 1 (um) para as primeiras N frases e 0 para as demais, onde N é um limiar sugerido para o número de sentenças.

Já a referência de Fattah e Ren (2009) segue o mesmo princípio apresentado por Satoshi e seus pares, além de assumir que as primeiras sentenças de um parágrafo são as mais importantes. As sentenças são ordenadas da seguinte maneira: a primeira frase no parágrafo recebe a pontuação 5/5, a segunda 4/5 e assim por diante, preterindo as sentenças que se localizam mais ao final do parágrafo.

Por fim, Barrera e Verma (2012) exploram o modelo de três posições. A primeira assume que as sentenças localizadas no início e no final do documento têm maior probabilidade de fornecer uma forte representatividade do conteúdo. A segunda prioriza apenas as partes que estão no início do texto. Já a última usa sentenças que estejam tratando do mesmo tópico de chamada do documento (similar ao tópico contido no título ou subtítulo) para criação do sumário.

Centralidade da sentença

Trata-se do vocabulário que é tido na intersecção entre uma e as demais sentenças de um dado documento (FATTAH; REN, 2009; ABUOBIEDA et al., 2012; KULKARNI; PRASAD, 2010). Esta técnica não utiliza tratamento semântico algum, limitando-se ao nível léxico. Um caminho para obter tal medida é utilizar algoritmos de similaridade de sentenças, como por exemplo, o proposto (*Bleu*) por Haque et al. (2010). A medida pode ser calculada da seguinte forma:

$$C\left(s\right) = \frac{K\left(s\right) \cap K\left(O_s\right)}{K\left(s\right) \cup K\left(O_s\right)} \tag{3.7}$$

onde, K é o conjunto com as palavras chave encontradas na sentença (s) e nas demais (Os).

Semelhança com o Título

Esta técnica trata-se do vocabulário coincidente entre uma sentença e o título do documento (SATOSHI et al., 2001; FATTAH; REN, 2009; KULKARNI; PRASAD, 2010; ABUOBIEDA et al., 2012). Neste caso, sentenças similares ao título, devem conter palavras que são consideradas importantes. Uma forma simples de calcular essa pontuação é a seguinte:

$$Pontuacao = \frac{Ntw}{T} \tag{3.8}$$

onde, Ntw é o número de palavras do título contidas na sentença e T é o número de palavras contidas no título.

3.3.3.3 Métodos de Pontuação Baseadas em Grafos

Nas abordagens baseadas em grafos, a pontuação é gerada pelo relacionamento entre as sentenças. Por exemplo, quando uma sentença se refere a outra, é gerado uma aresta com um peso associado entre as frases. Tais pesos são utilizados para gerar uma pontuação para cada sentença.

TextRank

Esta técnica trata-se de uma ordenação baseada no modelo de grafos para processamento textual (BARRERA; VERMA, 2012; MIHALCEA; TARAU, 2004). Consiste na extração de palavras-chave importantes do texto e então determina um coeficiente de "importância" para estes itens. Sentenças e relacionamentos que contenham um maior número destas palavras-chave obterão maiores pontuações.

Bushy path

A medida é definida simplesmente pelo número de arestas (*links*) conectando um nó (sentença) aos outros no grafo (FATTAH; REN, 2009).

Similaridade Agregada

Esta técnica gera uma medida de importância de uma sentença através da contagem do número de arestas conectando um nó do grafo (sentença) aos demais nós (*Bushy Path*) e, por fim, efetua um somatório dos coeficientes (similaridades) presentes nas arestas (FATTAH; REN, 2009).

3.3.4 Cenários de Avaliação

Serão aqui descritos os cenários de avaliação utilizados nos experimentos realizados.

3.3.4.1 Referência: Fluxo Padrão de Sumarização - Standard Summarization Flow (SSF)

A figura 4 mostra o Fluxo de Sumarização Padrão (SSF) que denota o processo clássico de sumarização extrativa (RADEV; HOVY; MCKEOWN, 2002; LLORET; PALOMAR, 2012). O SSF será usado como referência para a metodologia de avaliação adotada neste capítulo.

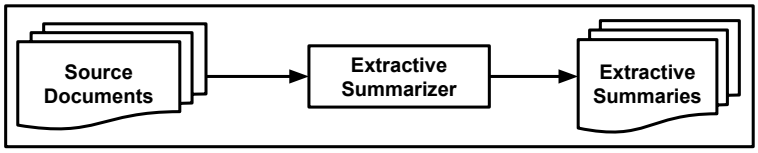


Figura 4 – Fluxo Padrão de Sumarização (SSF).

3.3.4.2 Pós-Processamento: Resolução de Expressões Anafóricas no Sumário - Anaphoric Expressions Solver in Summary (AESS)

A figura 5 exibe o cenário de aplicação do método AES (veja-se Seção 3.2) após aplicarem-se as técnicas ou sistemas de sumarização no corpus CNN. Como já apresentado, o objetivo principal do método de resolução de expressões anafóricas é abordar o problema de expressões anafóricas pronominais quebradas produzidas pelos métodos de sumarização extrativa. Portanto, aplicando-se o AES sobre os resumos gerados no cenário de referência, é possível eliminar todas as correferências pronominais quebradas nos resumos. Isso afeta diretamente a coesão dos resumos, pois muitos pronomes são resolvidos num passo de pós-processamento.

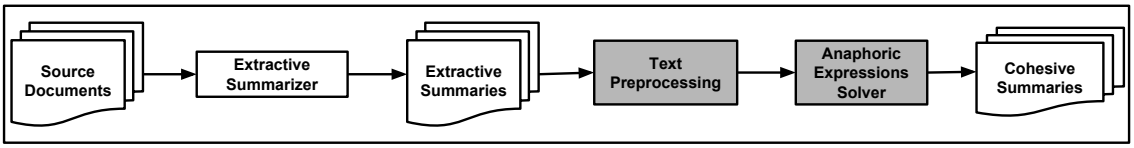


Figura 5 – Fluxo AESS.

3.3.4.3 Pré-Processamento: Resolução de Expressões Anafóricas no Corpus - Anaphoric Expressions Solver on Corpus (AESC)

A figura 6 mostra o último cenário de aplicação do método AES, que já foi descrito na Seção 3.3.4.3. A diferença básica entre esse cenário e o descrito na seção anterior é que as técnicas ou os sistemas de sumarização adotam uma nova versão do corpus CNN como entrada. O principal objetivo da aplicação do AES neste cenário é corrigir as menções que podem causar correferências quebradas nos resumos extrativos. Em outras palavras, o corpus de entrada é pré-processado antes de ser analisado pelo componente AES (figura 6) que finalmente gera documentos em um formato intermediário específico (etapa de pré-processamento) com as correferências tratadas no nível de sentença.

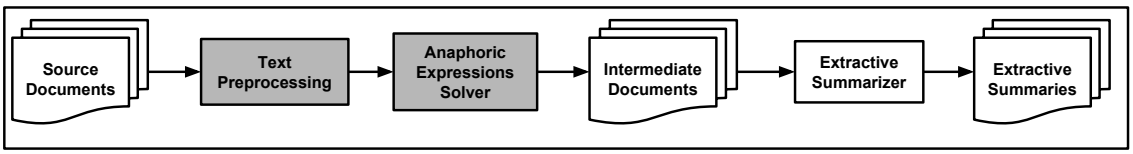


Figura 6 – Fluxo AESC.

3.3.5 Avaliação Quantitativa

Foram usadas duas medidas para avaliar o desempenho dos sistemas e técnicas de sumarização avaliadas nesta seção, como segue.

- **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) (LIN, 2004) mede a similaridade de conteúdo entre os sumários gerados pelos sistemas e os sumários de referência. A precisão, cobertura e a *F-measure* (BAEZA-YATES; RIBEIRO-NETO, 1999b) fornecidos pelo ROUGE foram usadas para realizar a avaliação quantitativa das técnicas propostas.

  A pontuação do *ROUGE-N* é calculada conforme apresentado na Equação 3.9:

  $$c_n = \frac{\sum_{c \in RSS} \sum_{gram_n \in C} Count_{match}(gram_n)}{\sum_{c \in RSS} \sum_{gram_n \in C} Count(gram_n)} \tag{3.9}$$

  onde *Count_match (gram_n)* é o número máximo de n-gramas que co-ocorrem em um resumo candidato e um resumo de referência (*Golden Standard*), enquanto $Count(gram_n)$ é o número de n-gramas no resumo de referência. Deve-se notar que a pontuação média de cobertura de n-gramas, C_n, é de fato uma medida de cobertura. A presente tese adotou o ROUGE-1 e o ROUGE-2.

- **Similaridade do Cosseno** (DONAWAY; DRUMMEY; MATHER, 2000): O grau de similaridade entre as sentenças do resumo gerado e o resumo de referência pode ser calculado usando similaridade do cosseno. Os termos (T) nas sentenças são ponderados usando Frequency–Inverse Document Frequency (TF-IDF) como na

Equação 3.12. A frequência do termo em um determinado documento é definido como o número de vezes que ele aparece, conforme mostrado na Equação 3.10:

$$TF_i = \frac{T_i}{\sum_{i=1}^{n} T_k} \tag{3.10}$$

onde T_i é o número de ocorrências do termo e T_k é a soma das ocorrências de todos os termos no documento. A *inverse document frequency* é uma medida de importância do termo:

$$IDF = log\left(\frac{N}{n_i}\right) \tag{3.11}$$

onde N é o número de sentenças no documento e n é o número de sentenças que contém o termo significativo. O peso correspondente é, portanto, computado como,

$$W_t = TF_i * IDF_i \tag{3.12}$$

O grau de similaridade entre sentenças pode ser medido usando a similaridade do cosseno como na Equação 3.13:

$$cos(X_i, Y_i) = \frac{x_i \cdot y_i}{||x_i| \cdot |y_i||} \tag{3.13}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{3.14}$$

onde X e Y são representações de um resumo candidato e um resumo de referência baseadas no modelo de espaço vetorial (STEINBERGER et al., 2007).

Todas as medidas mencionadas acima se concentram na informatividade do resumo extrativo. Tais características permitem a comparação direta entre os sistemas selecionados e as técnicas apresentadas na Seção 3.3.2.

Finalmente, como os resumos de referência compreendem cerca de 10% do tamanho total do corpus CNN, a mesma taxa de compressão foi usada em todos os experimentos relatados nas seções a seguir.

3.4 RESULTADOS EXPERIMENTAIS E DISCUSSÃO

O método proposto foi empregado em dois cenários de avaliação envolvendo todos os sistemas e técnicas de sumarização extrativa apresentados na Seção 3.3.2.

A metodologia de avaliação adotada aqui primeiro realiza uma análise quantitativa da razão das cadeias de referência espúrias e as expressões anafóricas inválidas produzidas pela etapa de pré-processamento usando o corpus CNN com 3.000 artigos de notícias (Seção 3.4.1). O principal objetivo aqui é filtrar apenas cadeias de referência contendo expressões anafóricas válidas nos resumos.

A Seção 3.4.3.1 apresenta a avaliação automática com uma análise comparativa envolvendo todos os cenários descritos nas seções 3.3.4.1, 3.3.4.2 e 3.3.4.3. Finalmente, a Seção 3.4.2 apresenta a avaliação humana usando o método proposto.

3.4.1 Resultados preliminares sobre a identificação de expressões anafóricas

Esta seção descreve os dois experimentos preliminares realizados, visando motivar o tratamento de expressões anafóricas para alcançar a coesão na sumarização extrativa.

3.4.1.1 Distribuição dos Pronomes Referenciais

Este experimento preliminar usa o *Stanford CoreNLP toolkit* para encontrar a distribuição dos pronomes referenciais no corpus CNN. Pronomes referencias ou anafóricos correspondem aos pronomes que fazem parte da mesma cadeia de correferência e que estão em sentenças distintas. Tal distribuição também pode ser considerada como uma estimativa aproximada do possível número de sentenças com correferências quebrados nos resumos gerados.

O corpus CNN contém 130.689 pronomes, o que corresponde a aproximadamente 5,69% de palavras no corpus, como mostrado na figura 3. O sistema de resolução de correferência CoreNLP identificou 115.791 pronomes fazendo parte de alguma cadeia de correferência. Esse número corresponde a 88,6% do total de pronomes do corpus CNN. Desse número, 82.675 pronomes são considerados referenciais, veja tabela 12.

Para obter resumos extrativos mais coesos, um crescente interesse tem sido focado na resolução da correferência pronominal (ORăSAN, 2009; STEINBERGER et al., 2007; SMITH; HENRIK; ARNE, 2012). Mais precisamente, para os pronomes que sua entidade referente aparece em uma sentença diferente. A tabela 12 mostra a distribuição das correferênicas pronominais categorizadas pelos principais tipos de pronomes.

Tabela 12 – Distribuição dos Pronomes Referenciais no Corpus CNN.

	Pessoal (Subj)	Objeto (Obj)	Possessivo (Poss)	Possessivo Adj. (Poss Adj)	Reflexivo (Reflex)
Total	42.600	10.773	7.701	20.489	1.112

De acordo com os resultados apresentados na tabela 12, os pronomes pessoais correspondem a 63,26% do número total de pronomes no corpus CNN. Curiosamente, quase a mesma taxa de pronomes pessoais fora relatada em Orăsan (2004) usando um corpus compreendendo 76 artigos científicos. Orăsan também apontou que apenas um terço dos pronomes tinham suas entidades de referência na mesma sentença, sugerindo que se uma sentença contendo um pronome fosse extraída, cuidados especiais precisariam ser tomados para evitar correferências pendentes. Em outro trabalho, Evans (2001) relatou que 67,94% dos pronomes de um corpus composto de documentos extraídos do SUSANNE (SAMPSON, 2002) e do BNC corpora (BURNARD, 1995) eram anafóricos.

Por um lado, tais resultados preliminares sugerem a necessidade de resolução pronominal na sumarização extrativa automática como um meio de evitar correferências quebradas nos resumos. Para controlar as repetições das entidades, algumas cadeias e menções não devem ser resolvidas ou substituídas por sua entidade de referência para evitar a reincidência da entidade representativa na mesma sentença, tanto no texto original quanto nos resumos.

3.4.1.2 Filtrando a Saída das Cadeias de Correferências Válidas

O principal objetivo deste experimento é avaliar a tarefa de filtrar as cadeias de correferências válidas para então extrair as entidades representativas e, por fim, aplicar o algoritmo AES proposto. Os resultados destes experimentos estão resumidos na tabela 13.

Tabela 13 – Avaliação final das cadeias e menções válidas.

	Cadeias	Cadeias Substituídas	Cadeias Eliminadas (%)	Menções	Menções Substituídas	Menções Eliminadas (%)
Total	49.329	24.511	50,31	115.791	45.149	61,01
Média/Doc	16,44	8,17	-	38,60	15,05	-

A análise dos textos sugere que muitas cadeias e menções não devem ser consideradas para a tarefa de sumarização. Por exemplo, a figura 7 mostra uma sentença processada pelo sistema de resolução de correferência de Stanford[5], em que o pronome *her* não será substituído por sua entidade representativa *(Angela Merkel)*, a fim de evitar a repetição dessa entidade na mesma sentença, de acordo com o método AES proposto.

Outro exemplo pode ser encontrado na figura 8, em que uma correferência entre dois pronomes é ignorada pela tarefa de sumarização no método AES.

Vale ressaltar que apenas um conjunto de todas as cadeias de correferência válidas foi usado nos cenários de avaliação experimental descritos no restante desta seção.

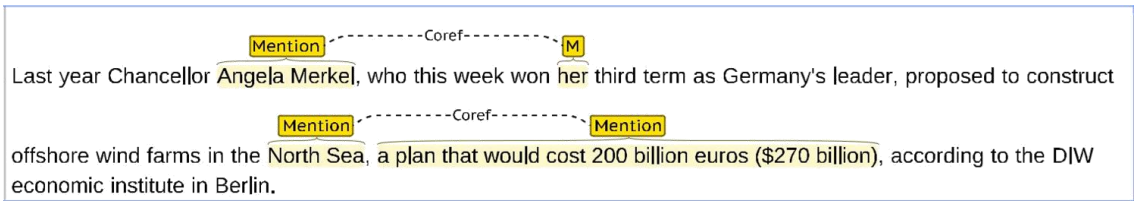


Figura 7 – Menções válidas pelo método AES.

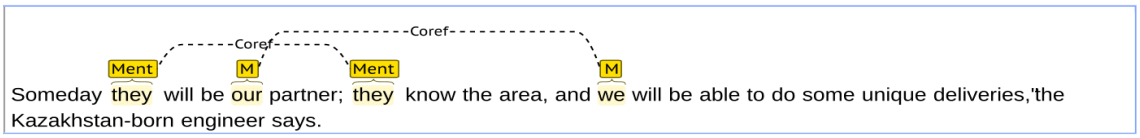


Figura 8 – Menções descartadas pelo método AES.

3.4.1.3 Análises de Correferências Quebradas

Os resultados deste experimento no corpus CNN estão resumidos nas tabelas 14 (sistemas) e 15 (técnicas), onde são mostrados:

- #Sumários: o número de resumos com correferências quebradas;

[5] http://nlp.stanford.edu:8080/corenlp/process

- #SentSummaries: o número de sentenças nos #sumários;
- #SentGolden: o número de sentenças nos resumos de referência;
- #Percentage(%): a razão entre #sumários e o número total de resumos no corpus CNN (3.000);
- #SentBrokenCoref: o número de sentenças com correferências quebradas nos #sumários;
- #Média: a média de sentenças com correferências quebradas por resumo.

Um aspecto importante a ser avaliado são as correferências quebradas presentes nos resumos gerados no cenário SSF. Para tanto, adotou-se a mesma metodologia de avaliação de Smith, Henrik e Arne (2012), a fim de avaliar o nível de coesão dos resumos, contando o número de correferências quebradas encontrados nos sumários. Além disso, a distribuição das correferências quebradas geradas pelos sumarizadores também fornece um desempenho de referência quanto ao nível de coesão dos resumos gerados pelos sistemas e técnicas avaliadas. Essa distribuição é resumida nas tabelas 14 e 15, que fornecem o número de correferências pronominais quebradas por resumo. Essas tabelas refletem o fato de que quanto maior o número de correferências quebradas nos resumos, menor é o nível de coesão. Segundo esse critério, o sistema C4J obteve o melhor desempenho entre os sistemas avaliados, pois selecionou menos sentenças com correferências quebradas (1.185 no total). O sistema AutoS vem na última posição, pois teve os piores resultados para todas as categorias de correferências pronominais avaliadas nesse experimento. Os resultados apresentados na tabela 14 mostram que mais da metade (56%) dos resumos gerados pelo AutoS tinham correferências quebradas, enquanto o C4J era o mais resiliente entre os sistemas avaliados. De fato, todos os resumos têm, em média, 46,50% de correferências quebradas, o que corresponde a uma taxa muito alta.

A tabela 15 mostra que as técnicas de sumarização geram muito menos problemas de coesão (26% em média). A técnica AS, em particular, gerou o maior número de resumos com baixa coesão, aproximadamente 32% do número total de resumos, enquanto a técnica SPT foi a menos afetada.

Os resultados apresentados na tabela 15 corroboram os resultados apresentados nos trabalhos anteriores (OUYANG et al., 2010; FERREIRA et al., 2013) que apontaram a importância da posição da sentença como um recurso para a sumarização extrativa, principalmente no domínio dos artigos de notícias. Uma possível explicação reside no fato de que, geralmente, os autores introduzem as entidades mais representativas no início do artigo. Então, as seguintes referências a tais entidades são substituídas por pronomes. Assim, a heurística simples de escolher as primeiras sentenças do texto tem menor probabilidade de gerar resumos com correferências quebradas em tal domínio. A técnica de PN alcançou o segundo

melhor resultado, possivelmente porque atribui pontuações mais altas às sentenças com nomes próprios.

Tabela 14 – Distribuição dos resumos com correferências quebrados por sistemas.

Sistemas	#Sumários	#SentSummaries	#SentGolden	%	#SentBrokenCoref	Média
AutoS	1.686	7.979	6.148	56,20	3.095	2,32
Aylien	1.490	7.161	5.482	49,67	2.761	1,85
HP-FS	1.219	6.120	4.701	40,63	2.152	1,77
C4J	1.185	6.929	4.377	39,50	2.136	1,80
Média	*1.395*	*7.047*	*5.177*	*46,50*	*2.536*	*1,82*

Tabela 15 – Distribuição dos resumos com correferências quebrados por técnicas.

Técnicas	#Sumários	#SentSummaries	#SentGolden	%	#SentBrokenCoref	Média
AS	954	3.894	3.484	31,80	2.074	2,17
N-GRAM	935	3.655	3.401	31,17	2.087	2,23
SC	931	3.774	3.369	31,03	2.003	2,15
BP	930	3.823	3.390	31,00	1.984	2,13
SL	923	3.842	3.378	30,77	2.029	2,20
TS	811	3.382	2.946	27,03	1.529	1,89
CP	809	3.474	2.927	26,97	1.561	1,93
BLEU	784	3.211	2.855	26,13	1.433	1,83
SPP	777	3.367	2.825	25,90	1.513	1,95
LS	768	3.299	2.800	25,60	1.427	1,86
TF/IDF	750	3.234	2.746	25,00	1.373	1,83
WF	737	3.236	2.697	24,57	1.348	1,83
UC	694	3.067	2.555	23,13	1.195	1,72
RT	672	3.153	2.498	22,40	1.241	1,85
ND	656	2.459	2.083	21,87	1.207	1,84
PN	650	2.837	2.410	21,67	1.080	1,66
SPT	564	2.882	2.377	18,80	972	1,72
Média	*785*	*3.329*	*2.867*	*26,17*	*1.532,71*	*1,93*

Considerando os resultados das tabelas 14 e 15, pode-se concluir que:

- Os resumos criados por todos os sistemas e técnicas (tabela 14) sofrem de problemas de coesão. A possível razão é que as melhores técnicas de sumarização são baseadas na frequência de palavras. Assim, a técnica individual ou a combinação aplicada não importa; os mesmos problemas de coesão surgirão.

- Se não houver nenhum tratamento adicional das correferências pronominais no momento da seleção das sentenças durante a geração do sumário ou alguma maneira de

lidar com as coreferências pronominais quebradas, pode-se sempre esperar problemas de coesão nos resumos gerados. Este fato justifica plenamente a solução proposta apresentada neste trabalho com o objetivo de eliminar as correferências pronominais quebradas nos resumos extrativos finais.

3.4.2 Avaliação Humana

A dificuldade em recorrer à avaliação humana tem sido um obstáculo considerável no desenvolvimento de sistemas automáticos de sumarização (SILVEIRA, 2015). Aqui serão analisados qualitativa e quantitativamente os resultados obtidos.

3.4.2.1 Avaliação Qualitativa

Na avaliação qualitativa da qualidade dos sumários gerados, existem vários problemas envolvidos, como custo, tempo, treinamento dos avaliadores, etc. Por um lado, a seleção dos aspectos linguísticos e textuais dos resumos a serem avaliados não é direta. Propriedades como coesão, fluência ou legibilidade são difíceis de definir objetivamente e sua avaliação difere de pessoa para pessoa e também depende de vários aspectos, como conhecimento prévio ou mesmo habilidades linguísticas. Além disso, o tamanho dos dados de entrada torna a avaliação ainda mais complexa e árdua. Ao avaliar a correção automática das correferências quebradas e a qualidade de um resumo, o texto original deve ser lido para verificar se a substituição automática do pronome por sua entidade representativa está correta.

Sabendo desses desafios, foram selecionados aleatoriamente 374 textos do corpus CNN cujos resumos apresentassem problemas de correferências quebrados. O método AES proposto sugeriu para cada pronome uma entidade para substituição automática. Adotando a plataforma *Amazon Mechanical Turk*[6], os avaliadores humanos foram responsáveis por avaliar a qualidade de cada resolução de correferência realizada por meio de questionários.

Cada pergunta do questionário é representada por uma *Human Intelligence Task* (HIT), na qual possui um resumo destacando um ou mais pronomes e entidades indicados pelo AES, um *link* para o texto original da notícia e duas questões relacionadas à (i) Legibilidade/Fluência Fluency - Claridade Referêncial (Reference clarity); (ii) qualidade textual; e (iii) coesão. A claridade referencial aborda as disfluências anafóricas relacionadas aos substantivos, pronomes e outras expressões referenciais não apropriadamente utilizadas. Estes aspectos são baseados nas características linguísticas propostas por Over, Dang e Harman (2007) e Steinberger e Jezek (2009). A seguir, um exemplo de um HIT analisado pelos avaliadores.

[6] https://requester.mturk.com/

She (Arianna Huffington) describes herself as a "sleep evangelist," has nap rooms in her offices at the AOL headquarters in New York and tries to start every day with meditation. Huffington, 62, founded Huffington Post in 2005, and two years ago sold it to AOL for $315 million.

Original text: <http://edition.cnn.com/2013/03/07/business/arianna-huffington-leading-women/>

1. Does the mention in parentheses correspond to their respective pronoun? (Yes or No)
2. After the replacement(s), has the summary become easier to read and understand? (Yes or No)

Os resumos foram avaliados por avaliadores humanos usando a plataforma *Amazon Mechanical Turk* onde pode-se determinar o perfil dos avaliadores, como nível de escolaridade, nível de confiança na plataforma, localidade, primeira língua, etc. Assim, apenas os avaliadores que possuem conhecimentos em processamento de linguagem natural e têm o inglês como primeira língua foram selecionados. Além disso, o nível de escolaridade escolhido foi o superior, com graduados e pós-graduados de diversas áreas. Os resultados dessa avaliação são relatados na seção a seguir.

3.4.2.2 Resultados

A avaliação humana foi conduzidada para avaliar a qualidade dos resumos gerados e fornecer suporte empiricamente para confirmar a hipótese de trabalho formulada: A resolução automática anafórica de um resumo extrativo melhora sua qualidade textual (coesão, fluência, legibilidade, etc.). Os resultados são mostrados na figura 9.

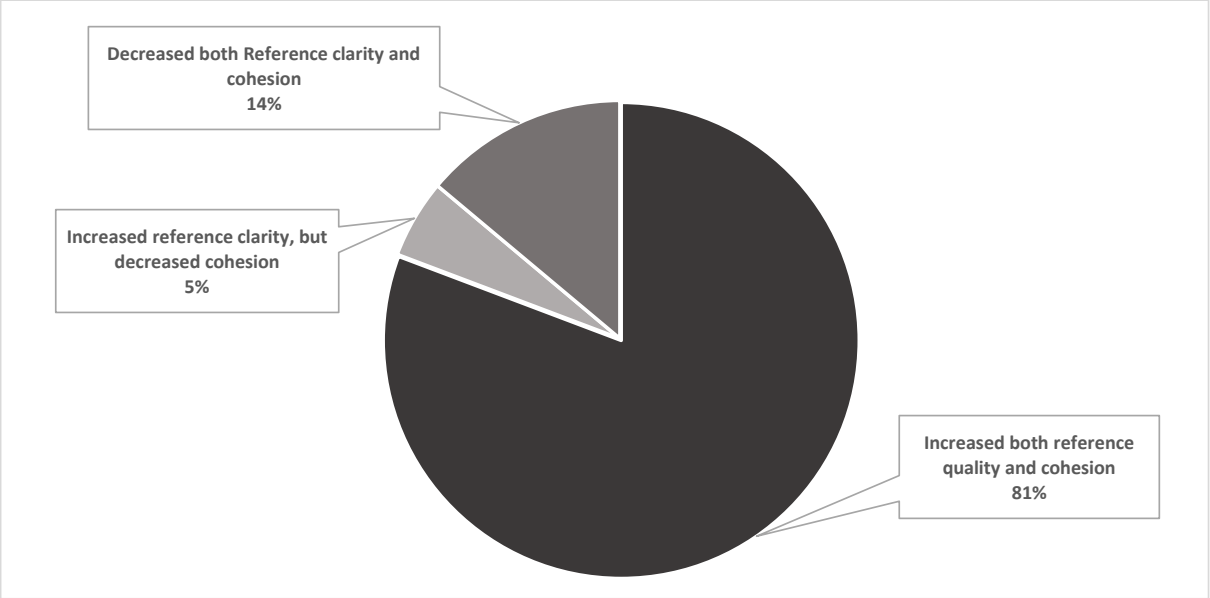


Figura 9 – Resultados da avaliação humana.

A análise da figura 9 permite observar que o método AES foi capaz de realizar a correta substituição das correferências, melhorando a coesão de 302 sumários, equivalente a 81% do total avaliado. Esta estatística significa que os avaliadores humanos deram respostas positivas às duas perguntas mostradas na Seção 3.4.2.1, ou seja, o método AES melhorou 81% dos resumos, tanto na clareza referêncial quanto na qualidade do texto.

A segunda avaliação analisou apenas os resultados da primeira questão "O substantivo entre parênteses corresponde ao seu respectivo pronome em particular? (*The subject in parentheses corresponding to their respective pronoun particularly?*)". Esta questão visa avaliar se a clareza referêncial dos resumos foi melhorada pelo método AES, ou seja, a menção sugerida entre parênteses pelo método AES é o substantivo correto referenciado pelo pronome. O método AES melhorou a clareza referêncial em 86% (322) dos resumos avaliados. Embora os sistemas de resolução de correferência de última geração não tenham apresentado elevado desempenho, o método AES foi capaz de extrair as cadeias de correferências mais assertivas do SCRS devido aos filtros empregados para remover cadeias de correferências espúrias.

Em 20 dos resumos (5%), embora o método AES tenha indicado a vinculação correta, os avaliadores humanos não consideraram que as substituições tivessem melhorado a coesão dos resumos. Cada um desses resumos foi analisado para identificar os possíveis problemas que ocorreram durante a resolução de correferência, que diminuíram a sua qualidade. A tabela 16 mostra fragmentos de resumos com tais problemas de coesão.

Como mostra a tabela 16, a entidade (*Rachel Canning*) indicada no fragmento [1] foi repetida na sentença. Esse problema ocorre porque o SCRS não foi capaz de identificar que a palavra "Canning" e a entidade "Rachel Canning" estão na mesma cadeia de correferência do pronome *she*. Assim, o método AES apontou que o pronome *she* estava desconectado

Tabela 16 – Fragmentos de resumos com problemas de qualidade de texto

[1] Canning alleged in her lawsuit that her parents forced her out of their home and that **she (Rachel Canning)** was unable to support herself financially.
[2] **She (a meat lover than a vegetable lover)** says Indians are developing new tastes because their incomes have grown – India's economy is slowly heading in the right direction again; people are earning more, they are traveling more and are being exposed to new, international cuisines.
[3] "I regret any hurt or anguish such comments may have caused any party and I look forward to greater understanding for peace and cooperation in future," **he (China's ambassador to Australia Monday)** wrote.
[4] In 1990, **he (a future Dr. Harold Freeman first envisioned in the 1980s)** pioneered the first-ever patient navigation program, training people from the community to listen and answer questions after a diagnosis.
[5] While laws in the West protect against discrimination, "it is kosher here in Asia to push youth and beauty," **he (Nok Air's Sarasin himself hedges on)** says.
[6] "He should stay on board that vessel until **He (The captain)** knows everybody is safely evacuated.
[7] "But we didn't know what was the maximum speed, so I thought it was normal," **he (One victim)** said.

(desvinculado). Uma possível melhoria para este problema é tratar as correferências nominais, tornando o método AES capaz de identificar a menção "Canning" referente à entidade "Rachel Canning".

Foi possível observar na análise dos fragmentos [2], [3], [4] e [5], que o SCRS incluiu várias palavras desnecessárias no conteúdo da MRE, o que levou a uma diminuição na qualidade desses resumos. No fragmento [6], o método AES realizou a substituição apenas na segunda ocorrência do pronome, uma vez que o SCRS não identificou o primeiro pronome na mesma cadeia de correferência. Finalmente, no fragmento [7], o método AES identificou corretamente a MRE e substituiu o pronome, mas de acordo com a avaliação humana, isso não melhorou a qualidade do resumo. Analisando a notícia original[7], foi visto que o autor não descreveu o nome da vítima. Embora a substituição tenha gerado um problema nominal de referência, não é possível identificar quem é a vítima.

Os avaliadores salientam que, em 52 resumos (14%), o método AES identificou erroneamente as entidades mais representativas e, assim, reduziu a coesão dos resumos. Isso aconteceu devido a erros nas cadeias de correferência geradas pelo SCRS. A figura 10 resume os resultados discutidos acima.

[7] http://edition.cnn.com/2013/07/26/world/europe/spain-train-crash/

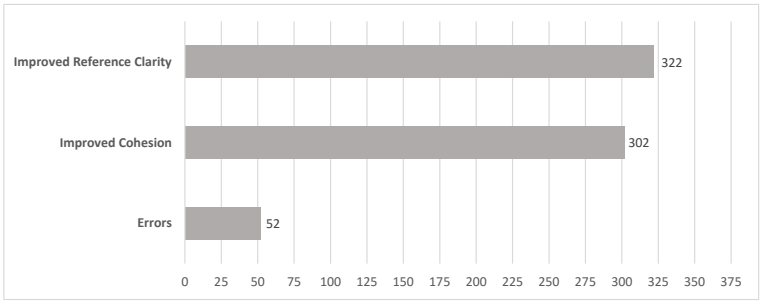


Figura 10 – Desempenho do Método AES.

3.4.3 Avaliação Automática

3.4.3.1 Avaliação Comparativa dos Cenários de Sumarização

Esta seção tem como objetivo realizar uma análise comparativa dos três cenários de avaliação descritos anteriormente. Para tanto, é introduzida a metodologia de testes estatísticos adotada neste trabalho, em seguida, os resultados comparativos são fornecidos como forma de verificar as hipóteses de trabalho levantadas.

Testes de significância estatística - A metodologia de avaliação utilizada baseia-se nos testes de significância estatística estruturados em três etapas:

1. O teste Shapiro e Wilk (1965) é usado para verificar se as pontuações (R1-F1, R2-F1 e CS) seguem uma distribuição normal;

2. Se o teste acima tiver um resultado positivo, o *t-test* pareado (GIBBONS; CHAKRABORTI, 2003) será aplicado; caso contrário, o teste de Wilcoxon (*Wilcoxon signed-rank test*) (GIBBONS; CHAKRABORTI, 2003) é usado;

3. Finalmente, o teste estatístico selecionado na etapa anterior é realizado duas vezes com nível de significância de 5% ($p - value < 0.05$): na primeira hipótese considera a região crítica bilateral; na segunda hipótese o teste é realizado assumindo a região crítica unilateal.

Essa metodologia de teste é usada para determinar se há uma diferença significativa no desempenho entre todos os cenários avaliados.

Comparação detalhada entre todos os cenários avaliados - tabelas 17 e 18 resumem os resultados da avaliação comparativa entre todos os três cenários de avaliação (SSF, AESS e AESC) apresentados nas seções 3.3.4.1-3.3.4.3. As pontuações (R1-F1 e CS) são fornecidas para os mesmos sistemas (tabela 17) e técnicas (tabela 18) discutidos anteriormente. Já as pontuações do R2-F1 são apresentadas na tabela 19 para os sistemas e tabela 20 para as técnicas.

Uma análise mais detalhada dos resultados da tabela 17 mostra que:

Tabela 17 – ROUGE-1 e CS. Avaliação comparativa de desempenho (%) e desvio padrão entre parênteses dos sistemas de sumarização. O desempenho geral mais alto é marcado em negrito e o grupo de cenários estatisticamente semelhante ao melhor desempenho é indicado por um †.

Sistemas	Referência				Método AES							
	Fluxo Padrão (SSF)				Resolução no Sumário (AESS)				Resolução no Corpus (AESC)			
	R	P	F1	CS	R	P	F1	CS	R	P	F1	CS
AutoS	49,87	47,02	48,40 (16,71)	64,62 (16,00)	59,26	40,59	**48,18** (14,70)	**66,68** (15,53)	55,62	38,75	45,68 (14,56)	64,00 (16,50)
Aylien	56,08	39,51	46,36 (15,61)	62,70 (16,78)	55,31	41,14	**47,18** (14,96)	**64,43** (15,97)	53,58	39,82	45,72 (14,73)	63,23 (16,28)
HP-UFPE FS	57,47	38,14	45,85 (15,67)	62,69 (16,70)	56,41	39,52	**46,48** (15,07)	**64,04** (16,07)	56,10	39,29	46,21† (15,15)	63,92† (16,47)
C4J	51,10	45,64	48,22† (16,10)	63,13 (16,26)	56,01	43,47	**48,95** (14,38)	**64,97** (15,43)	53,24	44,12	48,25† (15,29)	64,06 (16,26)

1. AESS alcançou os melhores resultados em termos de medida de CS entre todos os sistemas avaliados.

2. Entre o AESS e o AESC, o primeiro cenário de avaliação obteve uma melhora significativa em relação a R1-F1 e CS em comparação com o de referência. Em particular, a AESS foi capaz de melhorar o desempenho de referência dos sistemas AutoS, HP-UFPE FS e Aylien em relação a R1-F1.

3. Para o sistema de sumarização HP-UFPE FS, o AESS e o AESC obtiveram resultados estatisticamente semelhantes.

4. Finalmente, para o sistema C4J, nenhuma melhoria significativa foi alcançada para medida R1-F1 em qualquer um dos cenários testados pelo AES.

Uma análise comparativa mais ampla do desempenho dos três cenários de avaliação acima, usando testes de significância estatística, é fornecida na próxima seção.

De acordo com os resultados da tabela 18, as seguintes conclusões podem ser tiradas:

- Os resumos tratados pelo método AESS alcançaram as pontuações mais altas de F1 em onze das dezessete técnicas testadas.

- O método AESC obteve o segundo melhor resultado global sem diferença significativa no desempenho para doze técnicas sobre os resultados AESS, conforme a medida R1-F1.

- De acordo com os resultados dos sistemas mostrados na tabela 17, nenhum dos resultados das técnicas no cenário de referência (SSF) foi melhor do que os resultados do AES, considerando as pontuações de R1-F1 e CS;

- Em termos de medida de CS, os resultados são muito encorajadores para o cenário AESS, uma vez que teve uma diferença de desempenho significativa em comparação ao

Tabela 18 – ROUGE-1 and CS. Avaliação comparativa de desempenho (%) e desvio padrão entre parênteses dos sistemas de sumarização. O desempenho geral mais alto é marcado em negrito e o grupo de cenários estatisticamente semelhante ao melhor desempenho é indicado por um †.

Sistemas	Referência				Método AES							
	Fluxo Padrão (SSF)				Resolução no Sumário (AESS)				Resolução no Corpus (AESC)			
	R	P	F1	CS	R	P	F1	CS	R	P	F1	CS
AS	35,44	36,75	36,08 (12,44)	47,16 (16,77)	35,81	37,91	36,83† (12,24)	49,67† (16,38)	35,76	38,74	**37,19** (12,67)	**50,16** (16,43)
N-GRAM	41,08	35,92	38,33† (15,39)	50,08 (19,61)	41,34	37,33	39,23† (14,74)	52,67† (18,42)	41,39	37,98	**39,61** (15,44)	**52,79** (18,88)
SC	16,87	30,17	21,64 (11,23)	28,70 (18,65)	18,65	33,66	**24,00** (11,01)	**34,82** (17,47)	17,16	33,15	22,61 (10,99)	33,67 (17,70)
BP	35,04	36,36	35,69 (12,53)	47,26 (16,74)	35,49	37,61	36,52† (12,32)	49,98† (16,23)	35,88	38,85	**37,31** (13,15)	**50,28** (16,83)
SL	46,58	33,52	38,99 (14,64)	54,11 (18,46)	46,62	35,00	39,98† (14,12)	55,93† (17,51)	46,72	35,48	**40,33** (14,10)	**56,19** (17,18)
TS	44,28	37,04	40,34 (15,27)	53,81 (18,72)	44,11	38,55	**41,14** (14,60)	**55,90** (17,98)	42,96	38,51	40,61† (14,79)	55,51† (18,34)
CP	30,99	35,11	32,93 (12,72)	44,34 (18,22)	31,91	36,88	**34,22** (12,33)	47,36† (17,18)	31,68	36,77	34,04† (12,40)	**47,56** (16,97)
BLEU	22,77	36,15	27,94 (16,36)	34,34 (22,31)	24,13	38,82	**29,76** (15,53)	**40,46** (20,67)	22,85	38,40	28,65† (15,70)	38,84 (20,90)
SPP	30,78	35,38	32,92 (13,27)	44,65 (18,16)	31,54	37,11	**34,10** (13,01)	**47,46** (17,41)	31,18	36,85	33,78† (13,01)	47,07† (17,15)
LS	50,38	37,14	42,76 (15,71)	58,28 (18,21)	50,07	38,37	**43,45** (15,30)	**59,73** (17,82)	48,50	37,94	42,57† (15,20)	58,88† (17,92)
TF/IDF	53,08	36,84	43,49 (16,87)	59,04 (18,59)	52,59	38,15	**44,22** (16,21)	**60,57** (18,00)	51,86	37,61	43,60 (15,72)	60,33† (17,66)
WF	52,57	37,35	43,68 (16,07)	59,71 (17,84)	52,06	38,68	**44,39** (15,27)	**61,35** (17,16)	49,79	38,07	43,15 (15,21)	60,07 (17,75)
UC	44,30	35,72	39,55 (14,62)	53,58 (18,34)	44,45	37,06	**40,42** (14,14)	**55,72** (17,54)	43,85	36,13	39,61† (13,95)	54,66† (17,60)
RT	47,42	40,03	43,41 (14,10)	57,41 (17,47)	47,22	41,19	**44,00** (13,71)	**59,11** (16,83)	46,73	40,77	43,55† (14,01)	58,71† (17,08)
SPT	35,13	39,47	37,17 (15,18)	48,25 (20,10)	34,97	40,46	37,52† (14,59)	50,56† (18,96)	35,28	40,92	**37,89** (14,61)	**50,87** (19,08)
ND	38,11	37,18	37,64† (14,44)	49,68 (18,81)	38,06	38,59	**38,32** (13,75)	**52,15** (18,09)	37,76	38,20	37,98† (13,61)	51,86† (17,81)
PN	43,26	35,65	39,09 (14,38)	52,66 (17,86)	43,33	37,16	**40,01** (13,76)	**55,05** (17,02)	43,25	36,50	39,59† (13,86)	54,68† (17,36)

de referência. Além disso, para quatorze técnicas, o cenário AESC teve uma melhora significativa em relação ao de referência, apresentando resultados semelhantes aos da AESS.

- Com relação as pontuações de R1-F1, as técnicas de N-GRAM e ND são as únicas em que não há diferença significativa de desempenho entre os cenários do método AES e o de referência.

Em resumo, pode-se concluir que o desempenho geral de ambos os cenários de aplicação (AESS e AESC) da solução proposta para sumarização extrativa coesa supera o desempenho do cenário de referência em quase todas as técnicas de sumarização testadas em relação às tradicionais medidas: ROUGE-1 e similaridade do cosseno.

Tabela 19 – ROUGE-2. Avaliação comparativa de desempenho (%) e desvio padrão entre parênteses dos sistemas de sumarização. O desempenho geral mais alto é marcado em negrito e o grupo de cenários estatisticamente semelhante ao melhor desempenho é indicado por um †.

Sistemas	Referência			Método AES					
	Fluxo Padrão (SSF)			Resolução no Sumário (AESS)			Resolução no Corpus (AESC)		
	R	P	F1	R	P	F1	R	P	F1
AutoS	32,41	32,34	**32,38** (22,24)	38,31	26,88	31,59† (19,69)	34,06	24,20	28,30 (18,97)
Aylien	35,49	25,70	29,82† (20,57)	34,45	26,45	**29,92** (19,49)	32,29	24,90	28,11 (18,96)
HP-UFPE FS	36,60	25,33	**29,94** (20,42)	35,27	25,93	29,89† (19,36)	35,14	25,85	29,79† (19,36)
C4J	34,41	31,27	32,76† (21,52)	36,55	28,89	**32,27** (19,14)	34,73	29,31	31,79† (20,33)

Para os resultados das pontuações R2-F1 (tabela 19 e tabela 20), pode-se concluir que:

1. Os sistemas não apresentaram resultados estatisticamente relevantes para os cenários avaliados.

2. O cenário AESC apresentou resultados inferiores aos de referência para os sistemas AutoS e Aylien. O cenário AESS permanece estatisticamente igual à linha de base para todos os sistemas.

3. Os cenários AESS e AESC ultrapassaram o de referência para 8 técnicas. Para o cenário AESS, aquelas que ultrapassaram o de referência foram as técnicas SC, CP, SPP e UC, e para o cenário AESC aquelas que ultrapassaram o de referência foram as técnicas AS, N-GRAM, BP e SL. As outras técnicas não apresentaram resultados estatisticamente relevantes em relação ao de referência.

4. Em geral, os cenários AESS e AESC apresentam resultados estatisticamente semelhantes ou melhores que o de referência.

Comparação geral entre os cenários avaliados - A figura 11 resume as avaliações gerais de desempenho de todos os sistemas e técnicas de sumarizarição relatados pelas tabelas 17, 19, 18 e 20. Na figura 11, o intervalo entre 0,0 e 0,2 denota um valor de $p{-}value < 0.01$ e $p{-}value < 0.05$, respectivamente, ou seja, esse intervalo indica que os resultados são estatisticamente diferentes com nível de confiança de 95% e 99%, respectivamente, e são representados na figura 11 em um tom mais claro de cinza. Por outro lado, o valor de $p{-}value$ variando de 0,05 a 1,0 significa que não há diferença estatística entre os dois cenários, o que é representado usando uma tonalidade escura de cinza. Em todas as figuras, o cenário no eixo X é comparado ao cenário no eixo Y.

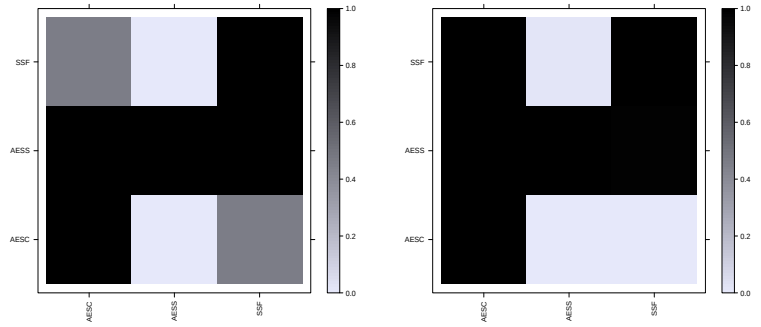
De acordo com a figura 11, AESS e AESC superam o de referência (SSF) em relação a R1-F1, R2-F1 e CS para a maioria dos sistemas e técnicas. De fato, o cenário AESS obteve as maiores pontuações gerais de R1-F1 e R2-F1, enquanto o AESC alcançou o

Tabela 20 – ROUGE-2. Avaliação comparativa do desempenho (%) e desvio padrão entre parênteses das técnicas. O desempenho geral mais alto é marcado em negrito e o grupo de cenários estatisticamente semelhante ao desempenho mais alto é indicado por um †.

Sistemas	Referência			Método AES					
	Fluxo Padrão (SSF)			Resolução no Sumário (AESS)			Resolução no Corpus (AESC)		
	R	P	F1	R	P	F1	R	P	F1
AS	15,34	16,77	16,02 (15,53)	15,59	17,49	16,48† (15,08)	15,76	18,44	**16,99** (15,70)
N-GRAM	20,32	18,90	19,58 (19,37)	20,17	19,70	19,93† (18,46)	20,64	20,68	**20,66** (19,29)
SC	7,82	12,38	9,58 (9,73)	7,65	13,91	**9,87** (9,98)	7,14	13,40	9,32 (9,90)
BP	15,11	16,12	15,60 (15,54)	15,42	16,92	16,13† (15,05)	16,25	19,03	**17,53** (16,09)
SL	24,66	18,11	20,88 (18,57)	24,16	18,65	21,05† (17,74)	24,26	19,40	**21,56** (18,01)
TS	24,53	20,96	**22,60** (19,72)	23,79	21,52	22,59† (18,68)	22,79	21,54	22,15† (18,67)
CP	12,72	15,34	13,90 (15,14)	13,12	16,22	**14,51** (14,67)	12,68	16,06	14,17† (14,90)
BLEU	12,36	19,36	**15,09** (17,00)	11,82	19,89	14,83† (16,39)	10,91	18,74	13,79 (15,97)
SPP	13,34	15,78	14,46 (15,38)	13,47	16,44	**14,81** (15,06)	13,05	16,26	14,48† (15,14)
LS	29,68	22,91	25,86† (20,61)	29,16	23,38	**25,95** (19,92)	27,47	22,67	24,84† (19,69)
TF/IDF	31,95	23,65	**27,18** (22,15)	31,01	24,06	27,10† (21,10)	29,73	23,21	26,07† (20,25)
WF	31,80	23,69	**27,15** (21,16)	30,82	24,12	27,06† (20,02)	28,14	23,17	25,41† (19,77)
UC	24,06	20,91	22,37 (18,79)	23,90	21,46	**22,61** (18,06)	22,51	20,14	21,26 (17,82)
RT	28,49	24,64	**26,43** (18,58)	27,62	24,98	26,24† (17,92)	27,13	24,81	25,92† (18,20)
SPT	19,72	22,48	**21,01** (18,71)	18,82	22,37	20,44† (17,66)	19,05	23,04	20,86† (17,83)
ND	19,88	19,90	19,89† (17,84)	19,45	20,56	**19,99** (16,96)	19,08	20,38	19,71† (16,76)
PN	22,56	19,92	21,16† (18,70)	22,57	20,78	**21,63** (17,74)	22,05	20,28	21,13† (17,98)

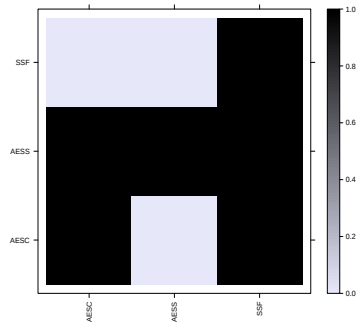
segundo melhor resultado geral, como pode ser visto nas figuras 11a e 11c. O cenário AESC perdeu para o de referência apenas nas pontuações R2-F1 para os sistemas, veja a figura 11b. Em termos de CS, o cenário AESC (figura 11d) produziu uma melhoria significativa em comparação com os outros cenários para as técnicas; enquanto que, para os sistemas (figura 11c), o cenário AESS foi o vencedor claro.

A conclusão é que os resultados gerais confirmaram a hipótese de trabalho formulada nesta tese de que uma análise profunda e a correção de expressões anafóricas envolvendo pronomes, seja no pré-processamento ou no pós-processamento dos resumos gerados, melhora a coesão do sumário e também aumenta o desempenho em termos das medidas tradicionais de avaliação de qualidade usadas pela comunidade de sumarização extrativa.

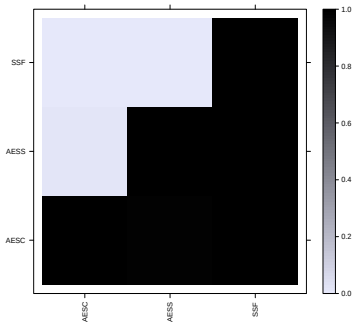


(a) Resultados do testes de significância dos sistemas (ROUGE-1 F-measure).

(b) Resultados do testes de significância dos sistemas (ROUGE-2 F-measure).



(c) Resultados do testes de significância dos sistemas (similaridade do cosseno) e técnicas (ROUGE-1/ROUGE-2 F-measure).



(d) Resultados do testes de significância das técnicas (similaridade do cosseno).

Figura 11 – Comparação geral dos sistemas e técnicas de sumarização nos três cenários de avaliação.

3.5 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este capítulo apresentou um novo método para a sumarização de textos extrativos que tenta produzir resumos mais coesos resolvendo expressões anafóricas pronominais. A implementação baseada em regras do método proposto é capaz de identificar e filtrar as cadeias de correferências espúrias do corpus de entrada (deixando intactas as entidades mais relevantes), e evitar correferências quebradas que possam dificultar a legibilidade dos sumários extrativos gerados. O método proposto amplia o trabalho relacionado, melhorando a coesão dos resumos gerados pelas abordagens clássicas de sumarização extrativa. O método foi extensivamente avaliado sob dois cenários distintos de aplicação, usando vários sistemas e técnicas de sumarização extrativa. O método AES alcançou resultados satisfatórios em avaliações quantitativas e qualitativas. No geral, os resultados obtidos com o corpus CNN demonstraram sua eficácia quando comparados ao cenário de sumarização de referência. Além disso, a arquitetura proposta para a sumarização

extrativa é independente do sumarizador utilizado.

O exaustivo número de avaliações para 17 técnicas e 4 sistemas foi necessário para analisar o impacto do método AES em cada técnica e sistema no processo de sumarização para os dois cenários avaliados. Além disso, mostra qual sistema possui o maior e menor número de resumos com correferências quebradas. Por exemplo, o método AES foi capaz de determinar corretamente que as técnicas de PN e SPT têm o menor número de correferências quebradas e, consequentemente, menos problemas de legibilidade e coesão. Como esperado, essas duas técnicas devem ter um número menor de correferências quebradas, uma vez que a primeira é baseada na seleção de sentenças com o maior número de nomes próprios e a segunda seleciona as primeiras ou a últimas sentenças do texto.

Apesar dos resultados encorajadores obtidos, ainda há espaço para muitas melhorias. O trabalho atual considera alternativas para gerar pronomes para evitar a repetição de entidades dentro de um resumo e o encurtamento de entidades repetidas, isto é, a troca da segunda ocorrência de uma entidade por uma referência mais curta ou um pronome. Além disso, acredita-se que um algoritmo de redução de sentenças permitiria a inserção de novas informações nos resumos, aumentando sua cobertura de informação. A viabilidade de explorar vários métodos de análise de informatividade, legibilidade e qualidade de texto será analisada para estimar a coesão dos resumos gerados automaticamente. Tal mecanismo pode permitir a seleção dos resumos mais coesos de um conjunto de vários resumos candidatos, independente de um resumo de referência. Essas contribuições serão exploradas nos próximos capítulos.

4 REINSERÇÃO DE PRONOMES EM SUMÁRIOS

Sistemas de Sumarização Automática (SAT) são exemplos de sistemas de processamento de linguagem natural que produzem saída em linguagem natural e, portanto, podem requerer algum tipo de módulo de geração de texto. O grau de sofisticação na geração do texto pode variar amplamente, mas dada a alta frequência de pronomes em textos de linguagem natural, é procedente esperar que um tratamento adequado de pronomes em sumarização possa levar à resumos de melhor qualidade. Pouca atenção tem sido dada à inserção (geração) de pronomes e o foco sempre foi na resolução de correferência, conforme abordado no capítulo anterior. A razão pode ser atribuída à falta de um bom *benchmark* para avaliação e/ou escassez de sistemas de geração de linguagem natural nos quais o módulo de inserção de pronomes pode ser conectado (KASHANI; POPOWICH, 2006).

A presente tese explora a SAT monodocumento, na qual visa identificar e extrair as informações mais relevantes a partir de um documento de entrada e apresentá-las de forma resumida (NENKOVA; MCKEOWN, 2012). Sistemas de sumarização podem auxiliar as pessoas, reduzindo o tempo gasto para identificar informações importantes a partir de uma coleção de documentos textuais. Nesse 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;

(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 esteja gramaticalmente correto, os resumos gerados também estarão corretos, já que eles são gerados utilizando as sentenças originais do documento transcritas *verbatim*. Contudo, resumos produzidos por abordagens extrativas, em geral, apresentam problemas de coesão, por exemplo, correferências em aberto quebram o fluxo de ideias entre as sentenças ou a repetição de entidades deixa o texto com uma leitura mais cansativa.

Atualmente, a maioria das abordagens propostas para sumarização monodocumento trata de aspectos relacionados à informatividade e à 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, 2015;

PARVEEN; STRUBE, 2015). Apesar de alguns trabalhos levarem em consideração a coesão nos resumos, as abordagens aplicadas, ainda são muito superficiais. Este capítulo propõe uma abordagem para melhorar a qualidade textual dos resumos extrativos, evitando a repetição de entidades (ex: nomes próprios) nos resumos. Explorou-se, nesta tese, essa questão com uma abordagem de inserção de pronomes que incorpora um sistema de resolução de correferência pronominal do estado da arte como parte do processo.

As principais contribuições deste capítulo são:

- Uma extensa investigação e avaliação de diversos sistemas de sumarização extrativa e monodocumento do estado da arte. Tal investigação complementa os sistemas e técnicas descritos nos capítulos 2 e 3, e foi conduzida utilizando os corpora do DUC 2001-2002 e o corpus CNN.
- Avaliação da importância da tarefa de inserção de pronomes para melhorar a fluência (qualidade textual) dos sumários.
- Uma abordagem para identificar e trocar entidades repetidas por pronomes nos sumários, gerando um novo sumário semi-extrativo e coeso.

O restante deste capítulo está organizado como segue: Na Seção 4.1 é apresentada a abordagem proposta. Na Seção 4.2 são apresentadas as configurações e os resultados dos experimentos realizados. Por fim, na Seção 4.3 são apresentadas as considerações finais do capítulo.

4.1 ABORDAGEM PROPOSTA

Uma visão geral da abordagem proposta é apresentada na figura 12, e as quatro etapas ilustradas são brevemente descritas a seguir:

1. **Sumarização:** Nesta etapa, passamos o documento de entrada para os sumarizadores automáticos do estado da arte, descritos na seção 4.2.2, para geração dos resumos extrativos. A abordagem proposta é independente de sistemas de sumarização, porém exige-se que os sumários sejam extrativos para que o método proposto consiga mapear as sentenças dos resumos em suas respectivas sentenças no documento original.
2. **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 Processamento de Linguagem Natural (PLN) executadas são: segmentação de sentenças, tokenização, lematização, atribuição das classes gramaticais, identificação de entidades nomeadas (NER), indentificação de gênero (feminino ou masculino), análise sintática e resolução de correferência.

3. **Identificação de Entidades Repetidas:** O método de identificação de entidades repetidas usa a mesma idéia de cadeias de correferências do algoritmo AES (proposto no capítulo 3) extraídas do Sistema de Resolução de Correferência de Stanford (SCRS) para mapear as entidades (ex: nomes próprios) em seus repectivos pronomes. Além do resumo extrativo, usa-se o documento original para fazer o mapeamento *entidade-pronome*. Para identificar se uma entidade está repetida no sumário, usam-se algumas heurísticas descritas na seção 4.1.1.

4. **Inserção de Pronomes** Nesta etapa, utilizam-se os mapeamentos *entidade-pronome* da etapa anterior e fazem-se as substituições automáticas nos resumos. No momento das substituições têm-se o cuidado de inserir os pronomes nas posições corretas, de modo que, esses pronomes não sejam conectatos por outras entidades do resumo. Ao final desta etapa, são gerados novos resumos semi-extrativos, a fim de melhorar a qualidade textual (legibilidade e coesão) do resumo. A seção 4.1.2 mostra algumas regras para inserção de pronomes e exemplos do sumário semi-extrativo.

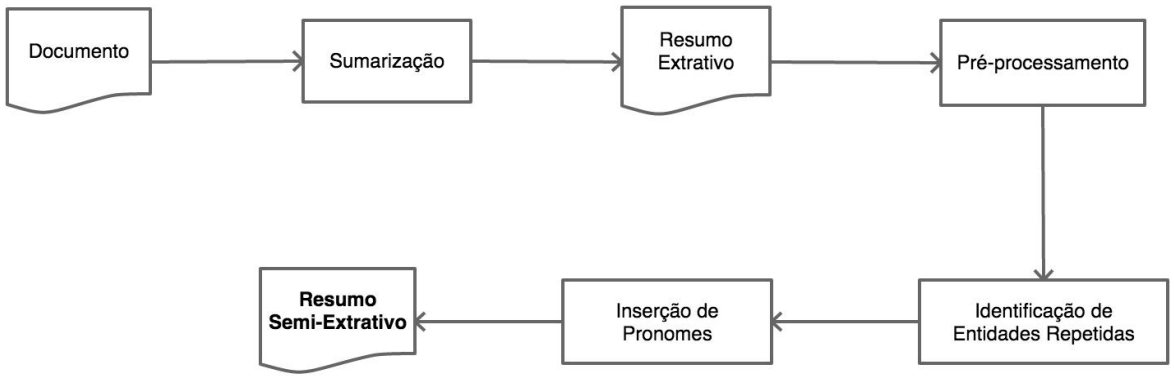


Figura 12 – Visão geral da abordagem proposta.

Como mencionando anteriormente, o SCRS pode encontrar todos os referentes para uma entidade específica, então, executando o SCRS uma vez no texto, teríamos uma cadeia de entidades, todas com o mesmo referente. Nosso objetivo é que as últimas entidades sejam sistematicamente substituídas por pronomes referentes às entidades anteriores. Para isso, o método proposto (IER&IP) pode ser dividido em duas fases: identificação de entidades repetidas e inserção de pronomes (substituição). Devido à sua importância para a abordagem proposta, as etapas 3, 4 são detalhadas na subseção a seguir.

4.1.1 Identificação de Entidades Repetidas

Para identificar as entidades repetidas, os resumos são pré-processados usando o CoreNLP. Durante esta fase são usadas as seguintes heurísticas para identificar e filtrar as entidades:

i Filtrar as cadeias de correferências pronominais do SCRS, deixando apenas pronomes e entidades que contêm ao menos um *token* com classe gramatical (*pos*) do tipo *NN, NNS, NNP, NNPS NNP, JJ, JJR, JJR ou JJS*;

ii Mapear as entidades nomeadas (NERs) do tipo *Person* no texto original e no resumo;

iii Aplicar os modelos do CoreNLP para identificação de gênero (MANNING et al., 2014) para entidades referidas mapeadas em NERs do tipo *Person*;

(a) Se o gênero não for automaticamente identificado pelos modelos do CoreNLP, são usados os pronomes (ex: he e she) das cadeias de correferências para identificar o gênero;

(b) Se foi possível identificar o gênero de uma entidade, o método não a marca como repetida. Diferentimente do trabalho Kashani e Popowich (2006) que marca como masculino o gênero das entidades não mapeadas. Segundo o autor, em artigos de notícias há maior predominância de entidades masculinas, porém essa abordagem pode gerar erros ao inserir pronomes com gênero inválido nos resumos. A estratégia proposta nesse capítulo tem como um dos objetivos minimizar os erros durante o processo automático de inserção de pronomes para que não afete a qualidade do texto;

iv Se uma entidade é identificada como repetida no resumo, mas o SCRS não conseguiu mapeá-la numa cadeia de correferência pronominal válida, então essa entidade é mantida no resumo final.

Os exemplos 4.1.1 e 4.1.2 a seguir mostram os mapeamentos (em negrito) da saída da etapa de identificação de entidades repetidas nos resumos. Esse mapeamento é repassado para a etapa de inserção de pronomes, descrita na próxima seção. No exemplo 4.1.1, o método proposto mapeou corretamente as entidades repetidas com seus respectivos pronomes no resumo. Perceba-se que a entidade *Diego Maradona* foi mapeada para o pronome *he*, já a entidade *Marcos Rojo* foi marcada como não aplicável (NA), pois não está repetida no sumário.

Exemplo 4.1.1 *(1) Football legend* ***Diego Maradona [he]*** *has reassured fans that* ***he*** *is fine after* ***he*** *was treated by doctors during Argentina's dramatic World Cup win over Nigeria on Tuesday.*

(2) ***Maradona [He]*** *was an animated presence during Argentina's 2-1 victory in St Petersburg, with pictures showing the 57-year-old celebrating wildly and directing a middle-finger salute at hecklers below following* ***Marcos Rojo [He/NA]****'s 86th-minute winner.*

(3) ***Maradona [He]*** *captained Argentina to World Cup triumph in 1986, scoring one of the greatest goals of all time in a quarterfinal victory over England as well as the infamous "Hand of God" goal.*

O exemplo 4.1.2 ilustra um mapeamento mais complexo durante a etapa de identificação de entidades repetidas. Diferentes tipos de pronomes foram mapeados, da mesma forma, algumas entidades foram marcadas como NA. Para entidades compostas, como *Prince Harry and Meghan Markle*, o método identifica o pronome de cada entidade para determinar seu gênero, e, por fim, identifica um pronome no plural que representa toda a entidade.

Exemplo 4.1.2 *(1) Britain's* ***- Prince Harry (he) and Meghan Markle (she) - [they]*** *have revealed further details about* ***their*** *wedding – including how the* ***bride [her/NA]****'s parents will be involved, the timing of their honeymoon and how* ***they*** *plan to remember* ***Harry [his]****'s mother,* ***Princess Diana [she/NA]****.*

(2) ***Meghan [She]****'s mother and father,* ***Doria Ragland [she/NA]*** *and* ***Thomas Markle [he/NA]****, who split up when* ***she*** *was young, will both attend.*

(3) *In the days before the wedding,* ***they*** *will fly into the UK to spend time with the British royal family, including* ***Queen Elizabeth II [she/NA]*** *and the* ***Duke [he/NA]*** *of Edinburgh, according to* ***Jason Knauf [he]****, communications secretary to* ***Prince Harry [he]****.*

(4) ***Meghan [She]*** *will spend* ***her*** *last night as a single woman at a hotel in an undisclosed location.*

(5) ***Meghan [She]*** *has opted not to have a maid of honor,* ***Knauf [he]*** *said.*

(6) ***Meghan (She) and Harry (he) [They]*** *will not depart for* ***their*** *honeymoon immediately after the wedding.*

(7) ***Knauf [He]*** *warned that Windsor is expected to be incredibly busy on the day, and recommends those hoping to stake out a spot on the processional route to leave themselves plenty of time.*

(8) ***Knauf [He]*** *said the ceremony will last an hour.*

4.1.2 Inserção de Pronomes

Nesta etapa é feita a substituição automática das entidades repetidas pelos seus repectivos pronomes. Para isso, recebe-se como entrada o mapeamento das entidades repetidas com seus repectivos pronomes e aplicam-se as seguintes regras:

i Para os casos gerais devem-se analisar as informações de gênero e usar pronomes pessoais do caso reto (ex: *he, she e they)*;

ii Se a entidade já é um pronome, nada é feito;

iii Se uma entidade é composta (ex: Prince Harry (he) and Meghan Markle (she) [they]), deve ser considerado, se existir, o pronome mais externo para substiuição (ex: they);

iv Se uma entidade estiver iniciando uma sentença, então a primeira letra do pronome correspondente deverá estar em maiúscula.

v Se uma entidade vier precedida de títulos de cortesia (ex. *Mr., Mrs., Ms. e Miss*) substui-se todo o trecho (entidade e título de cortersia) pelo pronome correspondente;

vi Se uma entidade for precedida por preposição ou verbo, devem ser aplicadas as seguintes regras;

(a) Se a maioria das preposições precede a entidade e não é seguida por *'s*, o pronome substituído deve ser pessoal do caso oblíquo (ex: *her, him e them*);

(b) Se a maioria das preposições precede a entidade e a entidade é seguida por *'s*, o pronome substituído deve ser possessivo (ex: *His, Hers, Theirs*);

(c) Se um verbo precede uma entidade (na sua forma base, passado, gerúndio, particípio ou presente) e a entidade não é seguida por *'s*, o pronome substituído deve ser pessoal do caso oblíquo (ex: *her, him e them*);

(d) Se um verbo precede uma entidade (na sua forma base, passado, gerúndio, particípio ou presente) e a entidade é seguida por *'s*, o pronome substituído deve ser possessivo (ex: *His, Hers, Theirs*).

Os exemplos 4.1.3 e 4.1.4 ilustram a análise de gênero dos pronomes durante o processo de substituição. No exemplo 4.1.3, as entidades Pedro e Mônica são de gêneros distintos, então o método proposto faz as substuições a partir da segunda ocorrência das entidades Pedro e Mônica sem que haja captura do pronome pela entidade errada. O resumo desse exemplo gerado após a etapa de inserção de pronomes é: *Pedro and Mônica are boyfriends. He went on the show with her. He did not like the show, and she loved it.*

Exemplo 4.1.3
Pedro [He$_1$] *and* ***Mônica [she$_2$]*** *are boyfriends.* ***Pedro [He$_1$]*** *went on the show with* ***Mônica [her$_2$]****.* ***Pedro [He$_1$]*** *did not like the show, and* ***Mônica [she$_2$]*** *loved it.*

No exemplo 4.1.4, as entidades Amanda e Amália são do mesmo gênero (feminino) então qualquer substituição poderá causar uma captura do pronome pela entidade não correspondente, afetando a legibilidade do resumo. Portanto, o método não faz as inserções de pronomes nesses casos, mantendo o resumo na forma original.

Exemplo 4.1.4
Amanda [She$_1$] *and* ***Amelia [she$_2$]*** *are childhood friends.* ***Amanda [she$_1$]*** *invited* ***Amelia [her$_2$]*** *to watch a movie.*

Os resumos finais dos exemplos mostrados na seção anterior, após a etapa de inserção de pronomes, são mostrados nos exemplos 4.1.5 e 4.1.6. Perceba-se que as entidades anotadas com NA não foram substituídas e algumas entidades foram mantidades devido à distância do seu referente.

Exemplo 4.1.5
*Football legend **Diego Maradona** has reassured fans that **he** is fine after **he** was treated by doctors during Argentina's dramatic World Cup win over Nigeria on Tuesday. **He** was an animated presence during Argentina's 2-1 victory in St Petersburg, with pictures showing the 57-year-old celebrating wildly and directing a middle-finger salute at hecklers below following **Marcos Rojo**'s 86th-minute winner. **Maradona** captained Argentina to World Cup triumph in 1986, scoring one of the greatest goals of all time in a quarterfinal victory over England as well as the infamous "Hand of God" goal.*

Exemplo 4.1.6
*Britain's **Prince Harry and Meghan Markle** have revealed further details about **their** wedding – including how the **bride**'s parents will be involved, the timing of their honeymoon and how **they** plan to remember **Harry**'s mother, **Princess Diana**. **Meghan**'s mother and father, **Doria Ragland** and **Thomas Markle**, who split up when **she** was young, will both attend. In the days before the wedding, **they** will fly into the UK to spend time with the British royal family, including **Queen Elizabeth II** and the **Duke** of Edinburgh, according to **Jason Knauf**, communications secretary to **Prince Harry**. **Meghan** will spend **her** last night as a single woman at a hotel in an undisclosed location. **She** has opted not to have a maid of honor, **Knauf** said. **They** will not depart for **their** honeymoon immediately after the wedding. **Knauf** warned that Windsor is expected to be incredibly busy on the day, and recommends those hoping to stake out a spot on the processional route to leave themselves plenty of time. **He** said the ceremony will last an hour.*

Ao final desta etapa, esperam-se sumários semi-extrativos, quando o método proposto for aplicável, com o menor número possível de entidades repetidas, melhorando a qualidade textual do resumo. Para validar a hipótese da abordagem proposta, a seguir são explorados diversos experimentos usando sistemas de sumarização do estado da arte.

4.2 EXPERIMENTOS

Nesta seção, são apresentados e discutidos os resultados dos experimentos conduzidos para avaliar diferentes aspectos da abordagem proposta. Dois experimentos foram executados abordando as seguintes questões: (i) Avaliação dos sistemas de sumarização extrativa (Subseção 4.2.2); e (ii) Análise do impacto da geração de pronomes na tarefa de sumarização (Subseção 4.2.3).

4.2.1 Configurações dos Experimentos

Todos os experimentos realizados utilizaram os corpora do DUC 2001, DUC 2002 e CNN. Na tabela 21 são apresentadas algumas estatísticas básicas para estes corpora.

Tabela 21 – 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** (LIN, 2004): As medidas de cobertura (R), precisão (P) e f-measure (F1) 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: $-m - fA$. 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): Esta medida avalia o grau de 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 100 palavras.

4.2.2 Avaliação dos Sistemas de Sumarização

Além da avaliação do impacto da geração de pronomes no processo de sumarização extrativa para melhorar a legibilidade e coesão dos resumos gerados, esse capítulo propôs-se a estender o número de sistemas e técnicas de sumarização extrativas descritas nos capítulos 2 e 3.

4.2.2.1 Sistemas Avaliados

Os sistemas do DUC 2001-2002 foram avaliados apenas nos corporas do DUC, pois eles não estão disponíveis para uso e sua identificação é oculta.

Para todos os corporas (CNN e DUC 2001-2002), selecionaram-se os quatro melhores sistemas (HP-FS, C4J, AutoS e Aylien) e técnicas (WF, TF/IDF, LS, RT) descritos nos

capítulos 2 e 3. Além disso, acrescentaram-se novos sistemas de sumarização do estado da arte (PLI baseda em conceitos + GE, Regressão Linear + PLI, MEAD, SUMMA, ALMUS, TextRank, LexRank, KL-Sum, G-Flow e SumBasic). Os sistemas TextRank, LexRank, KL-Sum e SumBasic foram implementados com base na biblioteca sumy [1].

Todos os sistemas selecionados são descritos brevemente a seguir:

Abordagem Baseada em Conceitos Utilizando PLI e Grafo de Entidade

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 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.

Oliveira et al. (2016b), evidenciou o bom desempenho na tarefa de sumarização 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, 2015; PARVEEN; STRUBE, 2015; DURRETT; BERG-KIRKPATRICK; KLEIN, 2016). Esses trabalhos tratam a tarefa de seleção de sentenças 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 nessa lacuna, Oliveira et al. (2016b) apresentou 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

[1] https://pypi.org/project/sumy/

do resumo a ser gerado; e (ii) Para garantir que sentenças redundantes não sejam incluídas nos resumos, utilizou-se a mesma estratégia 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 (GE) (GUINAUDEAU; STRUBE, 2013) para estimar a coesão local do resumo gerado.

Abordagem baseada em Regressão Linear Usando PLI

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, Oliveira et al. (2017) analisou os resultados obtidos nesse trabalho e em outros trabalhos (HONG; MARCUS; NENKOVA, 2015; WAN et al., 2015), e percebeu que essa conclusão podia 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; DEWALIYA; 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; DEWALIYA; 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 ordenação (*ranking*) 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 Oliveira et al. (2017). Dessa forma, é possível realizar uma melhor análise de cada resumo do que ordená-los ou combiná-los. Segundo Oliveira e seus colaboradores, a abordagem mais similar a sua é o sistema SumCombine introduzido por Hong, Marcus e Nenkova (2015) para a sumarização multidocumento. Porém, nenhum trabalho que tenha adotado algoritmos de regressão para estimar a informatividade de um resumo na tarefa

de sumarização monodocumento foi encontrado.

A referência Oliveira et al. (2017) listou duas limitações importantes na abordagem proposta por Hong, Marcus e Nenkova (2015), sendo elas:

1. Quantidade muito grande de sumários canditados, resultando num elevado custo computacional;

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. Outros aspectos importantes para a SAT, como posição e centralidade, não são considerados.

A abordagem proposta por Oliveira et al. (2017) visa suprir as limitações supracitadas do método proposto por Hong, Marcus e Nenkova (2015), além de estendê-la para a sumarização monodocumento. Nesse contexto, a abordagem basea-se 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.

Para geração dos resumos candidatos, 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 por Oliveira et al. (2017). 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.

MEAD

O MEAD (RADEV, 2001), é um sumarizador multilíngue superficial, escrito em Perl, bastante utilizado na área. Foi desenvolvido na Universidade de Michigan em 2000 e no início de 2001. Esse sumarizador implementa vários algoritmos de compressão, como os métodos baseados em posição, centroid [RJB00], TF*IDF e baseados em consulta. Entre os trabalhos relacionados à sumarização mono e multidocumento, Radev et. al, (2000) apresentam uma proposta de sumarizador e descrevem duas novas técnicas de avaliação para a sumarização mono e multidocumento baseadas na utilidade relativa (Radev & al. 2003) e redundância das sentenças. O sumarizador MEAD realiza a sumarização de conjuntos de documentos agrupados automaticamente por um sistema de detecção de

tópicos. Utilizando um conjunto de informações dos centróides desses conjuntos, o sistema MEAD seleciona as sentenças que provavelmente melhor descrevam o seu tópico. A análise de redundância é feita através de um algoritmo que verifica a similaridade entre sentenças, penalizando sentenças muito semelhantes à outras com maior métrica de relevância - isso é usado no contexto da sumarização multidocumento, em que se espera que informações repetidas apareçam de diferentes fontes.

SUMMA

A plataforma SUMMA oferece um conjunto de ferramentas de sumarização de texto (monodocumento, multidocumento e multi-idioma) que segue os preceitos arquitetônicos do *framework General Architecture for Text Engineering* (GATE). Essa plataforma foi desenvolvido por por Horacio Saggion (SAGGION, 2014) na Universidade de Sheffield para fornecer um resumo genérico de documentos.

GATE é uma plataforma para o desenvolvimento e implantação de tecnologias de processamento de linguagem em larga escala (CUNNINGHAM et al., 2002). Ele fornece três tipos de recursos: Recursos de Linguagem (LRs) que coletivamente se referem a dados; Recursos de Processamento (PRs) que são usados para se referir a algoritmos; e recursos de visualização (VRs), que representam componentes de visualização e edição. O GATE pode ser usado para processar documentos em diferentes formatos, incluindo texto simples, HTML, XML, RTF e SGML. Os documentos no GATE contêm um ou mais conjuntos de anotações. As anotações geralmente são atualizadas por PRs durante o processamento de texto. Cada anotação pertence a um conjunto de anotações e possui um tipo, um par de *offsets* (o intervalo de texto que se deseja anotar) e um conjunto de características (*features*) e valores usados para codificar vários tipos de informações. As características (ou nomes de atributos) geralmente são *strings*. Atributos e valores podem ser especificados em um esquema de anotação que facilita a validação e a entrada durante a anotação manual. O acesso programático aos conjuntos de anotações, anotações, características e valores é possível por meio da *Application program interface* (API) do GATE. Alguns componentes padrão do GATE são: tokenização, separador de sentenças, *Part-of-Speech Tagging* e reconhecimento de entidade nomeada.

O SUMMA usa alguns dos componentes padrão do GATE e também usa o documento para armazenar valores computados (criação de *features* e valores, e anotações especiais). Os componentes de sumarização podem ser combinados pelo usuário em um aplicativo customizado. SUMMA é um sistema baseado na combinação de um subconjunto de recursos (por exemplo, posição, título, frequência), o que permite a criação de resumos genéricos. O objetivo é fornecer uma ferramenta adaptável para o desenvolvimento, teste e implementação de soluções de sumarização customizadas. Os PRs calculam *features* numéricas para cada sentença no documento de entrada, o que indica quão relevante é a informação na sentença para *feature*. Os valores computados são combinados em uma

fórmula linear para obter uma pontuação para cada sentença que é usada como base para a seleção das sentenças para o resumo.

ALMUS

Almus é um sumarizador criado por (STEINBERGER et al., 2007), que gera um resumo a partir de um conjunto de documentos relacionados a um tópico. A abordagem de sumarização é baseada na análise semântica latente.

A versão utilizada neste capítulo sofreu algumas alterações da versão original descrita na seção 2.1. Essas modificações permitiram melhorar o desempenho do Almus.

Seguem as principais contribuições para melhoria no desempenho do Almus:

1. Uso do CoreNLP para pré-processamento das sentenças do documento original;
2. Uso dos termos do título das notícias (no corpus CNN) ou primeira sentença do documento (para os corporas DUC 2001-2002) com peso dois para montar a matriz usada na decomposição em valores singulares (SVD);
3. Para o corpus CNN foi usada a taxa de compressão no nivel de sentença, já para os corporas do DUC foi usada o número de palavras.

TextRank

O TextRank (MIHALCEA; TARAU, 2004) é um algoritmo baseado em grafos muito usado para a sumarização monodocumento e extração de palavras chave. Esse algoritmo usa o tradicional algoritmo Pagerank (BRIN; PAGE, 1998) e explora a ideia de representar as sentenças do documento como vértices em um grafo e arestas que 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.

G-Flow

O G-Flow é um sistema de sumarização extrativa automática proposto por Christensen (2014) e desenvolvido no Centro de Turing da Universidade de Washington, que busca equilibrar a coerência e a saliência. Para isso, introduz um modelo comum para a seleção e ordenação que equilibra coerência e relevância. A representação central do G-Flow é um grafo que aproxima as relações discursivas através de sentenças baseadas em indicadores, incluindo pistas de discurso, nomes deverbais, correferência, etc. Esse grafo permite estimar a coerência de um sumário candidato.

Para pré-processamento das sentenças dos documentos de entrada, o G-Flow usa o Stanford Core-NLP[2] e o Ollie[3].

[2] https://stanfordnlp.github.io/CoreNLP/
[3] http://knowitall.github.io/ollie/

LexRank

LexRank é um algoritmo de sumarização baseado em grafo, desenvolvido por Erkan e Radev (2004). Este algoritmo propõe representar as correlações entre as sentenças por meio de um modelo de grafos. Os vértices do grafo representam sentenças e as arestas representam o peso que indica a similaridade entre as sentenças.

As sentenças mais relevantes são selecionadas de acordo com a centralidade do autovetor (*eigenvector centrality*) obtida através do bem conhecido algoritmo PageRank (PAGE et al., 1998). Um esforço de investigação paralelo tem sido dedicado à formalização da tarefa de sumarização como um problema de máxima cobertura com grande quantidade de restrições baseadas na relevância da sentença dentro de cada documento. Os trabalhos focados nas correlações entre palavras e sentenças são exemplos de que a pesquisa está caminhando para este rumo, almejando chegar num patamar de obter não só uma medida de avaliação que informe o quão representativo é o sumário gerado, mas também se ele faz sentido como um todo, fator mais complexo e ainda em aberto.

KL-Sum

KL-Sum é um algoritmo de sumarização, introduzido por Haghighi e Vanderwende (2009), que usa como critério para selecionar as principais sentenças para formação do resumo, a frequência das palavras. Esse algoritmo acrescenta sentenças a um resumo, desde que diminua a divergência *Kullback-Lieber* (KL) (KULLBACK; LEIBLER, 1951).

SumBasic

Um algoritmo de sumarização simples e eficaz, introduzido por Vanderwende et al. (2007), que se baseia na observação da frequência relativa das palavras para selecionar as melhores frases para compor o resumo.

No SumBasic, a cada sentença S é atribuída uma pontuação refletindo quantas palavras de alta frequência ela contém,

$$Score\left(S\right) = \sum_{\omega \in S} \frac{1}{|S|} P_D\left(\omega\right)$$

onde $P_D\left(\cdot\right)$ inicialmente reflete as probabilidades de unigramas obtidas da coleção de documentos D. Um resumo S é progressivamente construído adicionando a sentença de pontuação mais alta de acordo com a fórmula acima.

Para reduzir a redundância, as palavras da sentença selecionada são atualizadas $P_D^{new}\left(\omega\right) \propto P_D^{old}\left(\omega\right)^2$. As sentenças são selecionadas dessa maneira até que o limite de palavras do resumo seja atingido.

4.2.2.2 Avaliação dos Sumarizadores

Como avaliação inicial, na tabela 22 são apresentado os resultados da medida de Intersecção de Sentenças (IS) entre os resumos gerados e os resumos de referência no corpus CNN. O melhor sistema de sumarização é destacado em negrito.

Tabela 22 – Resultados comparativos do desempenho dos sistemas para a medida IS no corpus CNN. O sistema com melhor desempenho global é destacado em negrito.

Sistema	**#Sentenças**	**#IS**	**%**
PLI Baseda em Conceitos + GE	1.1678	**3.343**	**28,63**
Regressão Linear + PLI	1.1670	3.247	27,82
MEAD	1.1569	3.132	27,07
SUMMA	1.2152	3.236	26,63
ALMUS	1.1770	3.057	25,97
HP-FS	1.1019	2.842	25,79
WF	1.0204	2.594	25,42
C4J	1.1966	3.014	25,19
TF/IDF	1.0204	2.513	24,63
AutoS	1.2358	3.029	24,51
Aylien	1.1019	2.657	24,11
LS	1.0204	2.386	23,38
TextRank	1.1781	2.721	23,10
RT	1.0204	2.295	22,49
G-Flow	1.1572	2.230	19,27
LexRank	1.1781	2.130	18,08
KL-Sum	1.1781	1.971	16,73
SumBasic	1.1781	1.361	11,55

O sistema PLI Baseda em Conceitos + GE apresentou o melhor resultado, identificando 28,63% (3343) das 10.754 sentenças presentes nos resumos de referência. Com pouca diferença vem seguido pelos sistemas Regressão Linar (do mesmo autor do PLI Baseda em Conceitos + GE) com 27,82% (3.247), MEAD com 27,07% (3.132), SUMMA com 26,63% (3.236) e o ALMUS com 25,97% (3.057). Todos esses sistemas superam os avaliados nos capítulos 2 e 3 para a medida IS.

Os piores sistemas foram o SumBasic 11,55% (1.361), KL-Sum 16,73% (1.971) e LexRank 18,08% (2.130), repectivamente. O sistema G-Flow foi criado originalmente para sumarização multidocumento, porém adaptamos ele para sumarização monodocumento, colocando cada *cluster* com apenas um documento, e não se obteve um bom desempenho na medida IS. As tabelas 23 e 24 apresentam os resultados da avaliação dos sistemas com

base nas medidas de cobertura (R), precisão (P) e f-measure (F1) do ROUGE-1 (R-1) e ROUGE-2 (R-2), repectivamente.

No corpus CNN, o sistema PLI Baseda em Conceitos + GE obteve o melhor resultado baseado na medida F1 do R-1, seguido pelo sistema de Regressão Linear + PLI. Ambos os sistemas não apresentaram diferença estatística significativa. Comparando com os demais e esses sistemas observou-se uma diferença significativa ao nível de 95% de confiança.

O sistema de Regressão Linear + PLI apresentou o melhor resultado de cobertura, mas teve uma piora na precisão, influênciando no resultado de F1. Oliveira et al. (2017) considerou apenas a medida de cobertura para medir a informatividade dos sumários gerados, por isso ele apresentou Regressão Linear + PLI como melhor abordagem para sumarização extrativa.

Assim como na medida IS, os sistemas LexRank, KL-Sum, G-Flow e SumBasic apresentaram os piores resultados, e os sistemas PLI Baseda em Conceitos + GE, Regressão Linear + PLI, MEAD, SUMMA e o ALMUS os melhores, reforçando a consistência nas medidas de avaliação.

Tabela 23 – [CNN - ROUGE-1] Resultados comparativos (%) e desvio padrão entre parênteses da avaliação dos sistemas na tarefa de sumarização monodocumento. O sistema com melhor performance é destacado em negrito e o grupo de sistemas estatisticamente similar a ele ($p - valor > 0.05$), se existir, é indicado usando o símbolo †.

Sistema	**ROUGE-1**		
	R	P	F1
PLI Baseda em Conceitos + GE	57,54	48,31	**52,52** (17.70)
Regressão Linear + PLI	58,58	46,95	52,12 (17.73)
MEAD	51,88	50,34	51,10 (18.11)
SUMMA	54,36	48,13	51,05 (17.48)
ALMUS	47,71	52,20	49,86 (17.02)
AutoS	48,82	50,37	49,58 (17.03)
HP-FS	52,15	47,05	49,47 (17.20)
Aylien	52,08	46,76	49,28 (16.60)
C4J	46,96	51,80	49,27 (17.13)
TextRank	49,59	47,70	48,63 (16.58)
WF	48,57	46,78	47,66 (17.44)
TF/IDF	49,20	45,10	47,06 (17.78)
RT	42,84	50,72	46,45 (16.44)
LS	46,59	45,87	46,23 (17.19)
LexRank	41,42	48,10	44,51 (15.46)
KL-Sum	44,53	42,43	43,45 (15.90)
G-Flow	38,24	48,39	42,72 (15.96)
SumBasic	27,18	47,47	34,56 (13.10)

A tabela 24 apresenta os resultados dos sistemas com base na medida R-2 para o corpus CNN. É possível observar que existe uma alta correlação entre os melhores sistemas avaliados pelas três medidas de avaliação (R-1, R-2 e IS), reforçando ainda mais os bons resultados dos sistemas selecionadas.

Os sistemas C4J e HP-FS superaram o AutoS na medida R-2, mas esse grupo de sistemas ainda permaneceram no mesmo intervalo de posições junto com o sistema Aylien. O sistema PLI Baseda em Conceitos + GE permaneceu como apresentando o melhor desempenho e, comparando com os demais sistemas, houve diferença estatística significativa.

Tabela 24 – [CNN - ROUGE-2] Resultados comparativos (%) e desvio padrão entre parênteses da avaliação dos sistemas na tarefa de sumarização monodocumento. O sistema com melhor desempenho é destacado em negrito e o grupo de sistemas estatisticamente similar a ele ($p - valor > 0.05$), se existir, é indicado usando o símbolo †.

Sistema	**ROUGE-2**		
	R	P	F1
PLI Baseda em Conceitos + GE	41,08	36,24	**38,51** (23.22)
Regressão Linear + PLI	41,36	35,05	37,94† (23.20)
MEAD	37,05	37,86	37,45 (23.83)
SUMMA	38,38	35,95	37,13 (22.98)
ALMUS	33,01	37,75	35,22 (22.95)
C4J	32,82	37,41	34,96 (22.81)
HP-FS	35,77	34,04	34,88 (22.82)
AutoS	32,77	35,94	34,28 (22.64)
Aylien	34,54	32,76	33,63 (22.30)
TextRank	32,59	33,08	32,83 (22.25)
WF	31,95	32,95	32,44 (23.47)
TF/IDF	31,98	31,72	31,85 (23.85)
RT	28,34	35,89	31,67 (21.92)
LS	29,76	31,68	30,69 (23.17)
KL-Sum	26,08	27,32	26,69 (21.53)
LexRank	24,08	29,74	26,61 (21.00)
G-Flow	22,65	29,96	25,80 (21.28)
SumBasic	12,52	22,89	16,19 (15.78)

Os próximos experimentos comparam o desempenho dos sistemas selecionados 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 100 palavras nos corpora do DUC 2001 e 2002, para formar o resumo de saída;

(iii) Os avaliadores humanos do DUC 2001-2002; e

(iv) Os melhores sistemas e técnicas do capítulo 2 e 3.

Nas tabelas 25, 26, 27 e 28 são apresentados os resultados deste experimento em termos das medidas do R-1 e R-2, para os corporas DUC 2001-2002.

Tabela 25 – [DUC 2001 - ROUGE-1] Resultados comparativos (%) e desvio padrão entre parênteses da avaliação dos sistemas na tarefa de sumarização monodocumento. O sistema com melhor desempenho é destacado em negrito e o grupo de sistemas estatisticamente similar a ele ($p-valor > 0.05$), se existir, é indicado usando o símbolo †.

Sistema	**ROUGE-1**		
	R	P	F1
Baseline			
1	44,81	44,22	44,51 (10.31)
Sistemas do DUC 2001			
T	44,53	46,11	45,31 (9.78)
P	43,90	43,43	43,66 (9.53)
R	42,97	43,65	43,31 (10.22)
O	42,99	42,75	42,87 (10.83)
V	38,77	47,44	42,67 (10.47)
W	40,13	44,47	42,19 (15.68)
Q	42,33	41,79	42,06 (9.24)
Y	41,50	42,15	41,82 (9.20)
Z	40,08	41,42	40,74 (9.65)
X	39,83	41,51	40,65 (11.77)
S	39,37	41,79	40,55 (9.63)
Sistemas Selecionados			
Regressão Linear + PLI	46,37	46,05	**46,21** (9.66)
PLI Baseda em Conceitos + GE	45,32	45,38	45,35 (9.72)
C4J	45,44	44,59	45,01 (9.69)
SUMMA	43,54	44,34	43,94 (9.93)
MEAD	41,55	46,13	43,72 (10.64)
TextRank	43,45	43,53	43,49 (10.17)
Aylien	42,04	44,07	43,03 (9.91)
ALMUS	41,30	43,90	42,56 (8.31)
AutoS	49,58	37,26	42,54 (8.37)
G-Flow	42,40	42,14	42,27 (9.89)
RT	40,14	42,75	41,40 (8.41)
WF	42,40	39,95	41,14 (8.40)
TF/IDF	43,35	38,91	41,01 (8.63)
LS	41,58	40,31	40,94 (8.00)
LexRank	38,40	43,70	40,88 (9.80)
KL-Sum	37,41	39,71	38,53 (8.33)
HP-FS	37,28	39,18	38,21 (10.72)
SumBasic	29,83	44,69	35,78 (9.37)

No corpus do DUC 2001, o sistema Regressão Linear + PLI obeteve a melhor performance com base nas medidas R-1 e R-2, superando o sistema T, considerado o melhor sistema do DUC 2001. Os sistemas PLI Baseda em Conceitos + GE e C4J atingiram resultados estatistamente similares ao sistema T. Os demais sistemas selecionados apesarem de atingirem resultados inferiores ao melhor sistema do DUC, ainda assim demostraram competitivos em relação aos demais sistemas do DUC 2001.

Apenas o sistemas Regressão Linear + PLI, T, PLI Baseda em Conceitos + GE e C4J superaram o sistema de referência (baseline), o que mostra que as 100 primeiras palavras dos documentos do DUC 2001 já represetam um sumário muito expressivo. Já em relação aos sumários gerados por humanos, os sistemas de sumarização automáticos ainda estão com desempenho relativamente distantes.

Assim com no DUC 2001, no DUC 2002 (tabelas 27 e 28) os sistemas de Regressão Linear + PLI e PLI Baseda em Conceitos + GE atingiram os melhores resultados gerais, superando o melhor sistema (28) da competição do DUC 2002 para as medidas de R-1 e R-2. Vale ressaltar que o sistema 28 foi ultrapassado pelo 19 em relação a medida F1 do R-2, pois o 19 obteve uma melhor precisão do que o 28.

Além dos sistemas Regressão Linear + PLI, PLI Baseda em Conceitos + GE e 28, apenas os sistemas 19, C4J e SUMMA superam o baseline. Outro fato, é que os melhores sistemas avaliados no DUC 2001-2002 apresentaram uma diferença de mais de 20% em relação ao desempenho dos sumários gerados por humanos, isso mostra que o processo de sumarização automática têm muito à evoluir ainda.

Para todos os corporas avaliadores (CNN, DUC 2001-2002) foi possível perceber uma alta correlação de desempenho entre os sistemas avaliados para as medidas R-1, R-2 e IS. Além do mais, os resultados obtidos variam substancialmente de um sistema para outro, demonstrando que uma grande diversidade de resumos é gerada por eles.

Tabela 26 – [DUC 2001 - ROUGE-2] Resultados comparativos (%) e desvio padrão entre parênteses da avaliação dos sistemas na tarefa de sumarização monodocumento. O sistema com melhor performance é destacado em negrito e o grupo de sistemas estatisticamente similar a ele ($p-valor > 0.05$), se existir, é indicado usando o símbolo †.

Sistema	**ROUGE-2**		
	R	P	F1
Baseline			
1	20,03	19,77	19,90 (11.68)
Sistemas do DUC 2001			
T	20,27	20,94	20,60† (10.91)
V	17,07	20,99	18,83 (11.69)
P	18,39	18,21	18,30 (10.61)
R	18,06	18,42	18,24 (11.22)
O	17,89	17,75	17,82 (11.70)
W	16,70	18,12	17,38 (11.10)
Q	16,54	16,32	16,43 (9.53)
Y	15,98	16,19	16,09 (9.90)
S	15,34	16,33	15,82 (9.73)
X	14,62	15,25	14,93 (9.61)
Z	14,62	15,16	14,89 (10.46)
Sistemas Selecionados			
Regressão Linear + PLI	21,10	20,95	**21,02** (11.66)
PLI Baseda em Conceitos + GE	20,25	20,27	20,26 (11.51)
C4J	20,35	19,96	20,15 (11.41)
MEAD	18,54	20,94	19,67 (11.44)
SUMMA	18,80	19,01	18,90 (11.38)
TextRank	17,83	17,78	17,80 (11.44)
Aylien	17,01	17,79	17,39 (11.07)
RT	16,11	18,52	17,23 (8.64)
AutoS	19,71	14,93	16,99 (9.31)
ALMUS	16,39	17,46	16,91 (9.62)
G-Flow	16,53	16,48	16,51 (11.15)
TF/IDF	16,70	16,03	16,36 (9.14)
WF	16,25	16,40	16,32 (8.90)
LS	15,63	16,33	15,97 (8.34)
LexRank	14,57	16,29	15,38 (10.11)
KL-Sum	13,21	14,07	13,63 (8.44)
HP-FS	12,55	12,52	12,54 (9.77)
SumBasic	9,65	14,08	11,45 (8.18)

Tabela 27 – [DUC 2002 - ROUGE-1] Resultados comparativos (%) e desvio padrão entre parênteses da avaliação dos sistemas na tarefa de sumarização monodocumento. O sistema com melhor performance é destacado em negrito e o grupo de sistemas estatisticamente similar a ele ($p-valor > 0.05$), se existir, é indicado usando o símbolo †.

Sistema	**ROUGE-1**		
	R	P	F1
Baseline			
1	47,15	48,03	47,58 (9.19)
Sistemas do DUC 2002			
28	48,07	48,63	48,35 (8.80)
19	45,98	50,55	48,16 (9.07)
21	47,75	47,36	47,55 (8.90)
29	46,43	47,42	46,92 (8.61)
27	45,97	47,91	46,92 (9.50)
23	43,27	50,26	46,51 (10.52)
15	45,44	46,45	45,94 (9.18)
31	46,17	45,71	45,94 (8.77)
18	44,32	45,70	45,00 (9.61)
16	43,73	45,16	44,44 (9.35)
25	41,30	45,17	43,15 (9.37)
17	17,34	51,17	25,91 (15.40)
30	6,62	69,11	12,09 (5.33)
Sistemas Selecionados			
Regressão Linear + PLI	49,78	49,35	**49,56** (8.31)
PLI Baseda em Conceitos + GE	48,85	48,95	48,90 (8.37)
Classifier4J	48,20	47,68	47,94 (8.78)
SUMMA	47,19	48,17	47,67 (8.91)
TextRank	46,49	47,12	46,80 (9.03)
Aylien	45,34	48,04	46,65 (8.85)
G-Flow	46,15	46,59	46,37 (9.25)
HP-UFPE FS	47,32	45,37	46,32 (9.23)
MEAD	43,85	49,09	46,32 (9.10)
AutoSummarizer	45,46	46,09	45,78 (8.39)
ALMUS	43,78	47,47	45,55 (8.93)
LexRank	42,02	48,25	44,92 (9.35)
WF	38,74	46,95	42,45 (8.45)
TF/IDF	39,72	45,57	42,45 (8.51)
RT	35,86	51,34	42,23 (9.31)
LS	38,22	47,08	42,19 (8.53)
KL-Sum	38,98	42,97	40,88 (8.88)
SumBasic	31,82	48,27	38,36 (9.82)

Tabela 28 – [DUC 2002 - ROUGE-2] Resultados comparativos (%) e desvio padrão entre parênteses da avaliação dos sistemas na tarefa de sumarização monodocumento. O sistema com melhor performance é destacado em negrito e o grupo de sistemas estatisticamente similar a ele ($p-valor > 0.05$), se existir, é indicado usando o símbolo †.

Sistema	**ROUGE-2**		
	R	P	F1
Baseline			
1	22,24	22,65	22,44 (10.15)
Sistemas do DUC 2002			
19	22,05	24,27	23,11† (10.19)
28	22,88	23,16	23,02 (9.99)
21	22,29	22,10	22,20 (10.10)
23	20,54	23,85	22,07 (10.19)
27	21,31	22,21	21,75 (10.03)
29	21,32	21,79	21,55 (9.85)
31	20,30	20,10	20,20 (10.12)
15	19,99	20,41	20,20 (10.21)
18	19,41	19,98	19,69 (10.30)
16	18,49	19,11	18,79 (10.15)
25	16,74	18,21	17,44 (9.81)
17	6,50	18,10	9,57 (8.16)
30	3,35	39,47	6,17 (4.45)
Sistemas Selecionados			
Regressão Linear + PLI	23,92	23,71	**23,81** (9.70)
PLI Baseda em Conceitos + GE	23,30	23,33	23,32† (9.75)
Classifier4J	22,69	22,43	22,56 (9.95)
SUMMA	22,01	22,47	22,24 (9.95)
MEAD	20,24	22,74	21,41 (9.95)
TextRank	20,98	21,19	21,08 (10.02)
Aylien	20,35	21,57	20,95 (10.10)
HP-UFPE FS	21,30	20,43	20,86 (9.95)
G-Flow	20,12	20,38	20,25 (10.35)
AutoSummarizer	19,97	20,50	20,23 (9.59)
ALMUS	19,26	20,91	20,05 (9.35)
LexRank	17,93	20,47	19,12 (10.06)
RT	15,54	24,69	19,08 (8.81)
WF	15,75	20,60	17,85 (8.25)
LS	15,62	20,66	17,79 (8.21)
TF/IDF	16,00	19,67	17,65 (8.23)
KL-Sum	15,19	16,83	15,97 (9.20)
SumBasic	11,58	17,04	13,79 (8.81)

4.2.3 Avaliando o Impactado da Inserção de Pronomes nos Sumários

Este experimento avalia a performance do método proposto (IER&IP) para identificação de entidades repetidas e inserção de pronomes descrito na seção 4.1. Para isso, mostra o número de sumários que a abordagem proposta (descrita na seção 4.1) foi aplicada. Quando a abordagem proposta é aplicada, significa que o sumário tem ao menos uma entidade repetida (ex: nome próprio). Além disso, esse experimento mostra o percentual de penetração do método proposto para geração de pronomes por sistema e corpora (CNN, DUC 2001 e DUC 2002). Esses resultados são apresentados nas tabelas 29, 30 e 31.

O IER&IP visa melhorar a legibilidade e coesão dos sumários extrativos, gerando a partir deles, novos sumários semi-extrativos. Esse método faz uso do mesmo algoritmo do AES para geração e filtro de cadeias de correferências válidas descrito na seção 3.2.2 do capítulo 3. A partir das cadeias de correferências e as heurísticas descritas nesse capítulo, são substituídas as entidades repetidas nos sumários por pronomes, se houver necessidade.

Como descrito anteriormente, o AES e o IER&IP compartilham o mesmo algoritmo para gerar cadeias de correferências válidas, então nesse capítulo o autor não julgou necessário efetuar uma nova avaliação manual da proposta, visto que o AES apresentou resultados satistafórios na avaliação manual descrita na seção 3.4.2.2. Outro motivo para não repetir a tarefa de avaliação manual é que essa tarefa é muito árdua e existem vários problemas envolvidos, como custo, tempo, treinamento dos avaliadores, etc.

A partir dos resultados encorajadores das avaliações manuais do método AES, avaliou-se nesta seção o impacto do novo método proposto (IER&IP) para inserção de pronomes em sumários extrativos. A tabela 29 apresenta o número de sumários que foram aplicados a abordagem de inserção de pronomes pelo método IER&IP e o número de sumários que não sofreram alterações. Foram avaliados um total de 3 mil sumários do corpus CNN. O sistema AutoS obteve o maior número de sumários com entidades repetidas (1.413), consequentemente o maior percentual de penetração do método proposto (47%). Sentenças nos sumários com entidades repetidas podem até mesmo aumentar a informatividade calculada a partir de medidas automáticas de avaliação de sumários, tais como ROUGE e similaridade do cosseno, no entanto os sumários são menos legíveis e coesos.

Os quatros sistemas mais impactados pelo método proposto no corpus CNN foram: AutoS (47%), SUMMA (42%), TextRank (39%) e MEAD (38%), respectivamente. Já os sistemas que apresentaram o menor número de sumários com entidades repetidas, são: SumBasic, KL-Sum, LS e RT. Apesar desses sistemas apresentarem menor impacto do método AES, eles não atingiram bons resultados na avaliação das medidas IS e ROUGE. Isso ocorreu porque esses sistemas selecionaram sentenças menos descritivas e com menor número de entidades (por exemplo, nomes próprios e substantivos em geral), impactanto na informatividade.

Nos 308 resumos avaliados do DUC 2001, tabela 30, pode-se notar que, o sistema AutoS (38%) também foi o mais impactado e o KL-Sum (5%) o menos. Além do AutoS,

os sistemas G-Flow (34%), sistema P (33%) e TextRank (29%) foram os mais impactados. Dentre os sistemas do DUC, o melhor sistema da competição, sistema T, obteve o menor número de sumários com entidades repetidas (11%) e de maneira geral o sistema KL-Sum (5%) foi o menos impactado.

Na tabela 31 são apresentados os impactos do IER&IP nos sumários dos sistemas do DUC 2002 e dos sistemas selecionados nesse capítulo. O sistema 15 foi o mais impactado (37%), seguido dos sistemas G-Flow (34%), Almus (34%) e 27 (31%). Já o sistema 30 (1%) com pior resultado na alivação do DUC 2002 obeteve o menor número com de entidades repetidas, seguido dos sistemas KL-Sum (5%), SumBasic (13%) e TF/IDF (16%).

Os resultados obtidos nestes experimentos demonstram que:

(i) todos os sistemas geraram sumários com entidades repetidas, demostrando a necessidade da inserção de pronomes para melhorar a legibilidade e coesão dos sumários;

(ii) não tem uma relação direta entre as medidas de informatividade (ROUGE e IS) e a análise do número de entidades repetidas nos sumários, pois essa última avalia critérios de qualidade textual; e

(iii) geralmente os piores sistemas nas avaliações das medidas IS e ROUGE têm o menor número de entidades repetidas, pois no geral esses sistemas selecionam sentenças com menor relevância.

Tabela 29 – Resultados da avaliação do impacto da geração de pronomes pelo método IER&IP nos resumos gerados pelos sumarizadores extrativos no corpus CNN, ordenados pelos sistemas cujos sumários têm maior número de entidades repetidas.

Sistema	**IER&IP**	**Sem Alteração**	**%**
AutoS	1413	1587	47
SUMMA	1269	1731	42
TextRank	1171	1829	39
MEAD	1135	1865	38
G-Flow	1062	1938	35
Aylien	1014	1986	34
ALMUS	999	2001	33
LexRank	875	2125	29
C4J	844	2156	28
WF	820	2180	27
Regressão Linear + PLI	804	2196	27
PLI Baseda em Conceitos + GE	803	2197	27
HP-FS	802	2198	27
TF/IDF	706	2294	24
RT	702	2298	23
LS	686	2314	23
KL-Sum	367	2633	12
SumBasic	334	2666	11

Tabela 30 – Resultados da avaliação do impacto da geração de pronomes pelo método IER&IP nos resumos gerados pelos sumarizadores extrativos no corpus DUC 2001, ordenados pelos sistemas cujos sumários têm maior número de entidades repetidas.

Sistema	**IER&IP**	**Sem Alteração**	**%**
AutoS	117	191	38
G-Flow	106	202	34
P	100	204	33
TextRank	89	219	29
WF	82	225	27
PLI Baseda em Conceitos + GE	82	226	27
TF/IDF	80	227	26
Q	76	228	25
Aylien	75	233	24
SUMMA	75	233	24
R	73	231	24
ALMUS	72	236	23
O	69	234	23
Regressão Linear + PLI	70	238	23
HP-FS	64	244	21
LS	62	245	20
RT	61	246	20
Z	58	246	19
MEAD	58	250	19
Y	56	248	18
V	55	249	18
C4J	55	253	18
LexRank	54	254	18
X	50	243	17
W	43	233	16
S	44	261	14
SumBasic	33	275	11
T	32	271	11
KL-Sum	16	292	5

Tabela 31 – Resultados da avaliação do impacto da geração de pronomes pelo método IER&IP nos resumos gerados pelos sumarizadores extrativos no corpus DUC 2002, ordenados pelos sistemas cujos sumários têm maior número de entidades repetidas.

Sistema	**IER&IP**	**Sem Alteração**	**%**
15	196	337	37
G-Flow	182	351	34
ALMUS	181	352	34
27	166	364	31
SUMMA	165	368	31
TextRank	165	368	31
AutoS	156	377	29
31	155	378	29
PLI Baseda em Conceitos + GE	151	382	28
21	150	383	28
Regressão Linear + PLI	149	384	28
C4J	148	385	28
Aylien	146	387	27
23	137	389	26
LexRank	131	402	25
29	128	405	24
19	124	408	23
HP-FS	118	415	22
MEAD	118	415	22
WF	113	420	21
16	109	424	20
17	109	424	20
18	109	424	20
RT	107	426	20
28	98	435	18
LS	97	436	18
25	89	443	17
TF/IDF	86	447	16
SumBasic	70	463	13
KL-Sum	25	508	5
30	3	530	1

4.3 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo, foi apresentada uma extensa análise comparativa entre diversos sistemas de sumarização automática monoducumento no domínio de artigo de notícias, dos seguintes corporas: CNN, DUC 2001-2002. Esses sistemas foram avaliados pelas tradicionais medidas ROUGE-1, ROUGE-2 e IS.

Os resultados obtidos demostraram que todos os sistemas têm um número significativo de resumos com entidades repetidas, sendo que ao inserir um pronome de forma correta, poder-se-ia melhorar a legibilidade e coesão do resumo.

Com base nas medidas do ROUGE-1, ROUGE-2 e IS para o corpus CNN, e levando em consideração todos os experimentos realizados, podemos apontar que os quatro melhores sistemas de sumarização são: PLI Baseda em Conceitos + GE, Regressão Linear + PLI, MEAD e SUMMA.

No DUC 2001 com base na medida do ROUGE-1, temos: Regressão Linear + PLI, PLI Baseda em Conceitos + GE, T e C4J. Já na medida do ROUGE-2, foram: Regressão Linear + PLI, T, PLI Baseda em Conceitos + GE e C4J. No DUC 2002 para ROUGE-1: Regressão Linear + PLI, PLI Baseda em Conceitos + GE, 28 e 19; e para ROUGE-2: Regressão Linear + PLI, PLI Baseda em Conceitos + GE, 19 e 28.

De maneira geral, alguns dos sistemas selecionados superam os melhores sistemas do DUC 2001-2002. Analisando os resultados gerados por todos os 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 sistemas analisados consegue manter um alto desempenho para todos os documentos ou grupos de documentos.

Analisando os resumos dos sistemas avaliados, percebeu-se um número significativo de entidades repetidas em cada resumo. Portanto, além das avalições quantitativas dos sumarizadores, temos como resultados obtidos nesse capítulo a análise do impacto do método proposto de inserção de pronomes em sumários extrativos. Esses sistemas de sumarização procuram maximizar o nível de informatividade dos resumos, mas não desenvolvem abordagens profundas para tratar problemas de qualidade textual. Isso corrobora a necessidade de abordagens de pós-processamento para melhoria da qualidade dos resumos extrativos, gerando a partir deles novos resumos semi-extrativos.

Revisando a literatura, observou-se que a abordagem de resolução de correferências tem apresentado resultados interessantes para melhoria de qualidade textual para a sumarização multidocumento. No entando, no melhor do conhecimento do autor desta tese, nenhum trabalho do estado da arte explora esse tipo de abordagem descrita no capítulo anterior e neste capítulo com tanta profundidade para sumarização monodocumento.

No próximo capítulo, dando continuidade a abordagem de geração de sumários semi-extrativos, será apresentado a proposta para redução de sentenças para sumarização monodocumento. No processo de redução faz-se o uso de heurísticas que levam em consideração a árvore de dependêndia das sentenças.

5 REDUÇÃO DE SENTENÇAS

Assim como os capítulos anteriores, o presente capítulo têm como foco o pós-processamento do sumários, isto é, melhorar a qualidade textual e informatividade dos resumos extrativos e, em seguida, gerar novos resumos semi-extrativos. Até então, os capítulos anteriores têm focado nas tarefas de fluência (resolução de anáfora e inserção de pronomes). O presente capítulo, tem como objetivo tratar as tarefas de compressão de sumários com uma abordagem proposta para redução de sentenças.

O autor desta tese, assim como Silveira (2015), trata a redução de sentença como uma etapa de área de pesquisa mais ampla chamada simplificação de texto. Segundo Silveira (2015), a redução de sentenças é, de fato, uma forma específica de simplificação do texto. Pode ser definida como simplificação sintática aplicada a uma sentença de cada vez. Esta área de pesquisa também é referida na literatura como compressão de sentenças ou simplificação de sentenças (VANDERWENDE et al., 2007).

A simplificação de texto é uma área recente de pesquisa em processamento de linguagem natural. As abordagens mais comuns à simplificação do texto dependem da indução de regras obtidas a partir de corpora anotados ou do desenvolvimento manual de regras sintáticas que definem o que simplificar. A maior parte dos trabalhos feitos até agora tem sido focados em transformações sintáticas no nível da sentença, atingindo resultados interessantes.

O pós-processamento inicia após as sentenças que compõem o resumo serem selecionadas. A partir daí, as sentenças serão primeiramente reduzidas em tamanho, enquanto ainda buscam preservar e transmitir seu conteúdo principal (esse processo é chamado de redução de sentenças). Como essa etapa remove as palavras das frases, a taxa de compactação resumida provavelmente será comprometida, ou seja, o resumo pode ter menos palavras do que a configuração da taxa de compactação determina. Portanto, após o primeiro conjunto de sentenças ter sido reduzido, um novo conjunto de sentenças deve ser adicionado ao resumo até que o número de palavras definido pela taxa de compactação seja atingido novamente, essa etapa é conhecida como revisão de compactação. Esse novo conjunto de sentenças será posteriormente reduzido. As tarefas de compactação devem ser repetidas até que a taxa de compactação seja finalmente atingida.

As tarefas de redução definem dois procedimentos que procuram incluir no resumo o máximo possível de informações. Em primeiro lugar, sentenças mais curtas são construídas a partir das sentenças extraídas, por meio de um procedimento de redução de sentenças. Em seguida, o procedimento de revisão da redução garante que o resumo atenda à taxa de compactação solicitada.

As principais contribuições deste capítulo são:

1. Um método baseado em grafo para simplificar as representações das sentenças que

substituem os grafos de dependência por um mais simples, mantendo as entidades alvo nele. O objetivo é acelerar a fase de aprendizado em uma plataforma de extração de relações proposto por Lima et al. (2014), aplicando várias regras para simplificação de grafos que restringem o espaço de hipóteses para geração de regras de extração. O artigo referente a essa contribuição (*Transforming Graph-based Sentence Representations to Alleviate Overfitting in Relation Extraction*) encontra-se na referência Lima et al. (2014);

2. Um método baseado em regras para redução de sentenças em sumarização extrativa;

3. Uma estratégia para revisão da etapa de redução das sentenças do sumário para que novas sentenças possam ser incluídas, atendendo a taxa de compressão. Tal estratégia demonstrou ser eficaz para aumentar a informatividade dos resumos gerados;

4. Avaliação manual do método de redução de sentencas.

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 considerações finais deste capítulo.

5.1 ABORDAGEM PROPOSTA

Usar a redução de sentenças é um passo para gerar um novo resumo, em vez de extrair resumos do texto original (VANDERWENDE et al., 2007). Nosso objetivo não é criar um sistema de sumarização, na verdade consiste em propor um método que faça redução de sentenças de um dado resumo, diminuindo a redudância e mantendo a coesão, fluência e legibilidade. Além disso, que possibilite ao sumarizador produzir resumos com o máximo de conteúdo possível para satisfazer o usuário, dado uma taxa limite de compressão. Ao final das iterações, tem-se um resumo semi-extrativo.

Como as abordagens propostas nesta tese são focadas em sumarização monodocumento e extrativa, viu-se a redução de sentença (também conhecida como simplificação ou compressão de sentença) como um meio de criar mais espaço para capturar conteúdo importante.

Para apresentação da abordagem proposta, primeiramente são definidas várias regras para identificação das estruturas sintáticas (Subseção 5.1.1) e em seguida o algoritmo proposto para pós-processamento de resumos é detalhado (Subseção 5.1.2).

5.1.1 Regras de redução de sentenças

Há diversas heurísticas que podem identificar estruturas linguísticas que contenham informações redundantes sobre um conteúdo já expresso nas sentenças dos sumários. Vale ressaltar que a simplificação proposta neste trabalho é no nível de sentença. O método

proposto não olha a relação entre as sentenças. Ele procura manter as sentenças reduzidas com a mesma gramaticalidade, informatividade e fluência das sentenças originais.

Nessa tese, 15 regras são direcionadas para encontrar as estruturas liguísticas. Essas estruturas são direcionadas levando em consideração a árvore sintática de cada sentença, que é fornecida pelo analisador sintático. Essas estruturas são identificadas na árvore usando o CoreNLP. A seguir são descritas brevemente as regras:

Frases adjetivas delimitadas por pontuação (*adjective phrases delimited by punctuation*): Extrai frases adjetivas que são delimitadas por pontuação a partir da sentença de entrada.

> **Exemplo 5.1.1**
> *The Pendleton Round-Up, now 103 years old, plans to hire its first paid executive.*

Frases adverbiais delimitadas por pontuação (*adverb phrases delimited by punctuation*): Extrai frases adverbiais que são delimitadas por pontuação a partir da sentença de entrada.

> **Exemplo 5.1.2**
> *However, the man forgot where he has placed his keys.*

Frases preposicionais iniciais (*prepositional phrases at the beginning*): Remove frases preposicionais do início de uma sentença a partir da sentença de entrada.

> **Exemplo 5.1.3**
> *For a second time this month the Bulgarian parliament is to hold an extraordinary sitting.*
>
> *In 2012, he found and forged a new majority, turned weakness into opportunity and sought, amid great adversity, to create a more perfect union.*

Frases preposicionais delimitadas por pontuação (*prepositional phrases that are set off by punctuation*): Remove frases preposicionais que são definidas por pontuação a partir da sentença de entrada.

> **Exemplo 5.1.4**
> *Floyd Mayweather is open to fighting Amir Khan in the future, despite snubbing the Bolton-born boxer in favour of a May bout with Argentine Marcos Maidana, according to promoters Golden Boy.*

> *Paul, on Monday evening, walked to the primary school at the end of the block, during a very heavy thunderstorm.*

Frases preposicionais começando com "*according to*": Remove frases preposicionais começando com as atribuições "according to" da sentença de entrada.

> **Exemplo 5.1.5**
> *(i) Engineering students with bad handwriting may end up getting zero marks, according to a Mumbai University circular.*

Frases preposicionais começando com "*to*" (*specific prepositional phrases starting with "to"*): Remove frases preposicionais específicas começando com "*to*" (presumivelmente expressando uma intenção) da sentença de entrada.

> **Exemplo 5.1.6**
> *Governor Rick Scott declared a state of emergency for 26 counties to support emergency response operations for communities facing heavy rain and flooding.*
>
> *Kristen Stewart has enrolled at the University of California, Los Angeles to study English Literature.*

Sintagmas nominais iniciais (*lead noun phrases*): Extrai sintagmas nominais ou frases nominais parentéticas iniciais a partir da frase de entrada.

> **Exemplo 5.1.7**
> *Twenty years later, France defeated Croatia 4-2 to become World Cup champion for the second time in history.*

Cláusulas relativas não-restritivas: Remove cláusulas relativas não-restritivas da sentença de entrada. Essas cláusulas adiciona informação a sentença, mas não a limita, nem a define, a frase ficará com mesmo sentido se ela for retirada.

> **Exemplo 5.1.8**
> *Slovenian flag carrier Adria Airways has signed a code share agreement with Air Serbia, under which the Serbian airline will have the exclusive right for the Ljubljana-Belgrade route.*
>
> *Our school, which is over 300 years old, is getting a new auditorium.*

Sentenças com trechos entre parênteses (*constructs included in parentheses*): Remove construções incluídas entre parênteses (indicadas por tags PRN - *Parenthetical*) da sentença de entrada.

> **Exemplo 5.1.9**
> *John Smith has been appointed CKO (chief knowledge officer) of the merged company.*
>
> *Guido Cavalcanti (1255?-1300) had a profound influence on the writings of Dante.*

Sentenças com trechos entre colchetes (*bracketed content*): Remove o conteúdo entre colchetes da frase de entrada.

> **Exemplo 5.1.10**
> *Why can't we do the same thing [provide government-funded grants to independent filmmakers] in this country?*

Frases participativas delimitadas por vírgulas: Remove frases participativas que são delimitadas por vírgulas da sentença de entrada.

> **Exemplo 5.1.11**
> *The Nanticoke Junior High School girls won the 9th grade basketball championship, defeating Wyoming Area recently.*

Frases parentéticas: Remove frases que explicam ou detalham outras informações sendo expressas.

> **Exemplo 5.1.12**
> *My father, with his fear of crowds, did not come with us to the state fair.*

Frases parentéticas podem ser introduzidas por uma preposição (PP), ou um advérbio (ADV ou ADVP), ou uma conjunção (CONJ ou CONJP). Essas frases são delimitadas por símbolos de pontuação, como parênteses, vírgulas ou traços.

Apostos: Podem ser vistos como um tipo específico de expressões parentéticas, composto por frases nominais que descrevem, detalham ou modificam seu antecedente (também frases nominais).

Exemplo 5.1.13
José Sócrates, the Prime Minister, and Jaime Gama want to cut the salaries of their offices.

Apostos são identificadas por frases nominais (NP) ou frases adjetivas (AP) normalmente delimitadas por duas vírgulas ou dois traços.

Frases preposicionais no nível da sentença *Sentence-level prepositional phrases*: São frases encabeçadas por uma preposição usada para incluir informações adicionais na sentença de entrada.

Exemplo 5.1.14
Despite bad weather, the airplane arrived on time.

Frases preposicionais (PP) que serão direcionadas são introduzidas por uma preposição e são seguidas por uma sentença (S) que contém o sujeito e o verbo principal.

Cláusulas relativas apositiva *Appositive relative clauses*: São cláusulas que detalham as informações transmitidas por uma frase nominal.

Exemplo 5.1.15
My only brother Pedro, who is a Catholic priest, lives in Lima

Cláusulas relativas positivas são introduzidas por um pronome relativo e sua estrutura é definida por uma frase complementar (CP) que deve ser precedida por uma vírgula.

5.1.2 Algoritmo: simplificação de sentença baseada em sintaxe

A figura 13 ilustra o algoritmo proposto para remoção de trechos não importantes das sentenças, baseado nas regras descritas na seção anterior. O algoritmo recebe como entrada o resumo composto por uma ou mais sentenças, em seguida inicia a etapa de pré-processamento usando o Stanford Natural Language Processing Toolkit (CoreNLP)[1]. As tarefas de Processamento de Linguagem Natural (PLN) executadas no pré-processamento são: segmentação de sentenças, tokenização, lematização, atribuição das classes gramaticais, identificação de entidades nomeadas (NER), análise sintática e resolução de correferência.

Após a fase de pré-processamento, são aplicadas, em sequência, todas as regras de simplicação descritas na seção anterior. Ao final, é analisado se a taxa de compressão (número de palavras) do sumário é igual ao sumário original, se não, é recuperada a sentença seguinte na ordenação de relevância do sumário extrativo, que atenda a taxa de

[1] https://stanfordnlp.github.io/CoreNLP/

compressão e retorna ao fluxo inicial. Esse processo é repetido até que o limite da taxa de compressão seja atingido.

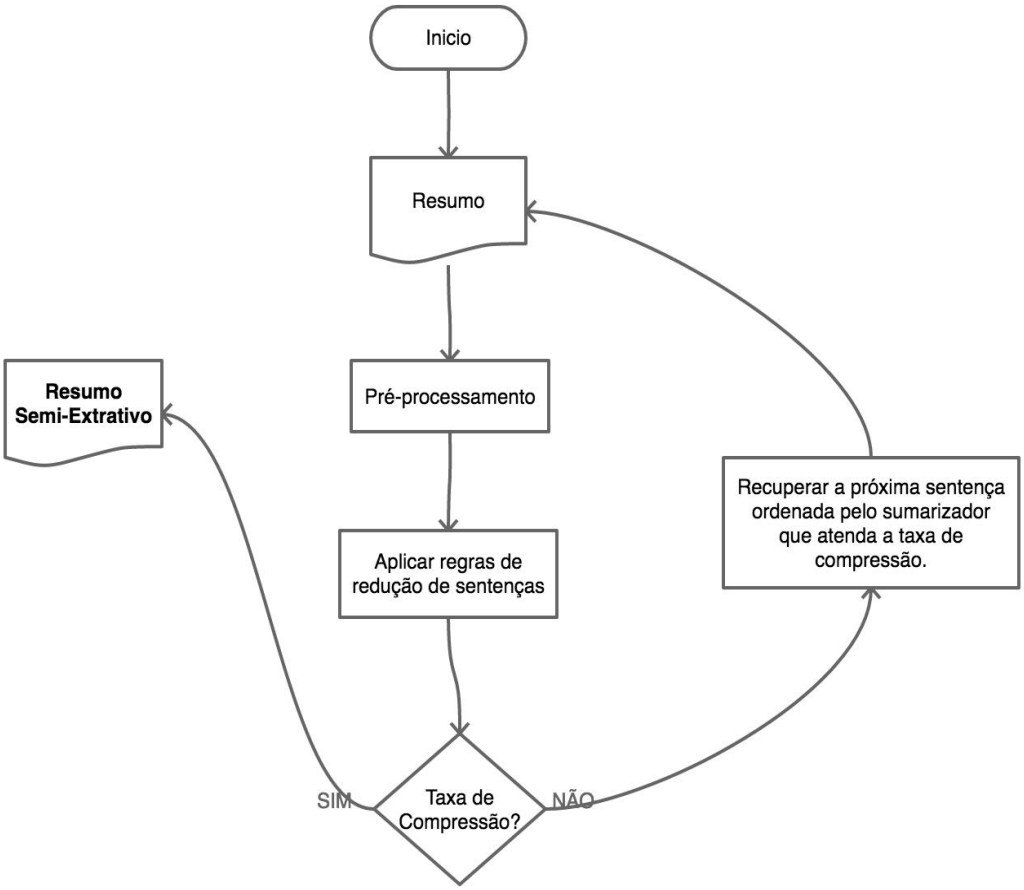


Figura 13 – Fluxograma do algoritmo proposto.

5.2 EXPERIMENTOS

Esta seção apresenta e discute os experimentos realizados para avaliar diferentes aspectos da abordagem proposta, nas tarefas de redução de sentenças para sumarização monodocumento semi-extrativa. Os experimentos foram conduzidos buscando avaliar manualmente os seguintes aspectos: (i) se a tarefa de compressão de sentenças mantém a legibilidade (também chamada de gramaticalidade e fluência) da sentença original; (ii) se a tarefa de compressão de sentenças mantém a informatividade (também chamada de importância e representatividade) da sentença original; e (iii) se o percentual de compressão é relevante o suficiente para adicionar novas sentenças e/ou palavras ao sumário, aumentando assim a sua informatividade.

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

5.2.1 Configurações dos Experimentos

Todos os experimentos foram realizados no conexto de redução de sentenças, para isso adotou-se um grande corpus[2] anotado para compressão de sentenças e uma ferramenta[3] para avaliação manual do método proposto nesta tese.

Corpus. *Google sentence-compression* (FILIPPOVA; ALTUN, 2013) é um corpus com 250 mil pares de sentenças extraídas de notícias em inglês e coletadas a partir do serviço Google News[4]. Esses pares de sentenças são formados pelo título (manchete) e a primeira sentença da notícia, pois são conhecidos como semanticamente semelhantes (DORR; ZAJIC; SCHWARTZ, 2003). Segundo Filippova e Altun (2013), pouquíssimos títulos são compressões extrativas da primeira sentença, portanto, simplesmente procurar por pares onde o título é uma subsequência das palavras da primeira sentença não resolveria o problema de obter uma grande quantidade de dados paralelos. É importante ressaltar que os títulos são sintaticamente bem diferentes das sentenças "normais". Por exemplo, eles podem não ter verbo principal, omitir determinantes e parecer incompletos.

Assim, em vez de usar o título original, Filippova e Altun (2013) usaram o título para encontrar uma compressão extrativa adequada da primeira sentença da notícia, combinando *lemas* de palavras (substantivos, verbos, adjetivos, advérbios) e identificadores das correferências das entidades do título com as da sentença. De modo que a árvore de dependência da compressão seja uma subárvore da árvore da sentença original. Esse corpus é usado principalmente para treinar sistemas supervisionados de compressão de sentenças.

Ferramenta de avaliação. Foi adotada a ferramenta *figure-eight*[5] para avaliação humana. Ela usa o conceito de *Human-in-the-loop* (HITL)[6] para criar e validar modelos de aprendizagem de máquina usando inteligência humana e de máquina. Em uma abordagem tradicional da HITL, as pessoas estão envolvidas em um círculo virtuoso no qual elas treinam, ajustam e testam um algoritmo específico.

Figure-eight é similiar a ferramenta Amazon Turk usada no capítulo 3. No atual capítulo Figure-eight foi adotada porque é mais usada por grandes empresas como a própria Amazon, Autodesk, Google, Facebook, Twitter, Cisco Systems, GitHub, Mozilla, VMware, eBay, Etsy, Toyota e American Express para ajudar a melhorar os modelos de todas as faixas, sejam classificadores de texto, algoritmos de visão computacional ou modelos de pesquisa e recuperação de informações. Pode-se criar grandes quantidades de dados de treinamento altamente precisos para um dado caso de uso, ajustar modelos com percepção humana e testá-los para garantir que as decisões sejam precisas e acionáveis.

Metodologia de avaliação. Foram selecionadas aleatoriamente 10 mil sentenças originais do corpus do Google, em seguida aplicou-se o método proposto baseado em

[2] https://github.com/google-research-datasets/sentence-compression
[3] https://www.figure-eight.com/
[4] https://news.google.com/?hl=en-US&gl=US&ceid=US:en
[5] https://www.figure-eight.com/
[6] https://www.figure-eight.com/resources/human-in-the-loop/

regras para redução de sentenças e, por fim, foram selecionados de forma aleatória as 256 sentenças com maiores percentuais de redução para avaliação da abordagem proposta.

Durante a etapa de avaliação humana aplicaram-se questionários para avaliar a fluência e a informatividade das novas sentenças reduzidas. Para garantir a qualidade do processo de avaliação foram adotados os seguintes critérios na ferramenta figure-eight:

1. permitir apenas avaliadores de 9 países (Estados Unidos, Canadá, Austrália, Reino Unido, Bahamas, Barbados, India, Jamaica e Nova Zelândia), cuja língua oficial é o inglês;

2. os avaliadores devem possuir uma acurácia mínima de 70%. Essa porcentagem é a precisão mínima que um avaliador deve manter durante a avaliação para continuar avaliando. Se o avaliador ficar abaixo dessa precisão a qualquer momento, ele ou ela será removido do trabalho e todas as suas respostas serão desconsideradas ou não confiáveis.

   Para calcular essa porcentagem, foram usadas algumas perguntas de teste com casos positivos e negativos em cada questionário. Sendo que para cada avaliação válida, o avaliador deveria responder no mínimo uma pergunta de teste;

3. um número de 3 avaliadores para cada questionário. No final é calculado um percentual de confiança (descrito na próxima seção);

4. o tempo mínimo por questionário deve ser de 25 segundos;

5. cada avaliador só pode responder no máximo 10 questionários. Sendo uma sentença avaliada por questionário; e

6. todos os avaliadores devem ser nível 2 na ferramenta. Esse nível indica maior qualidade, ou seja, consiste num grupo menor de colaboradores mais experientes e com maior precisão.

   A figure-eight possui 3 níveis de avaliadores, sendo o nível 3 os avaliadores mais qualificados na ferramenta e consequentemente as avaliações com maior preço.

Na seção a seguir descreve-se os detalhes dos questionários aplicados, a média resultante da pontuação de confiança dos 3 avaliadores e os resultados da avaliação intrínseca (Subseção 5.2.2.1)

5.2.2 Avaliação Humana do Método Proposto

Este experimento avalia manualmente a performance do método proposto para redução de sentenças em tarefas de sumarização extrativa. Para isso, adotou-se avaliação intríseca baseada em questionários. Um modelo de um questionário com uma sentença original e uma sentença reduzida para avaliação é apresentado abaixo. A primeira pergunta refere-se

se a sentença reduzida preserva a mesma informatividade da sentença original. Já a segunda refere-se se a sentença original pode ser substituida pela reduzida, preservando a mesma fluência.

Original Sentence: Vilaça was the first to declare that “Jones Bule” (as he called the Englishman) had made of Ramalhete “a veritable museum”.

Simplified Sentence: Vilaça was the first to declare that “Jones Bule” had made of Ramalhete “a veritable museum”.

1. Does the simplified sentence preserve the same idea of the original one?
2. Can the simplified sentence replace the automatic one without losing information?

Cada questionário foi avaliado por três avaliadores. A tabela 32 mostra algumas estatísticas básicas do processo de avaliação. Foram 523 canditados a avaliador, dos quais, seguindo as restrições descritas na seção anterior, apenas 177 avaliadores foram aprovados com um total de 1.677 avaliações. Essas avaliações incluem as perguntas de testes para validação das avaliações, das quais foram 768 avaliações válidas das 256 sentenças e 909 avaliações dos questionários de testes.

Dos nove países selecionados para que permita avaliações, quatro tiveram avaliadores. A figura 14 mostra a distribuição do número de avaliadores válidos (alta confiança) por País. A Índia ficou em primeiro, seguido dos Estados Unidos, Canadá e Grã-Bretanha.

Tabela 32 – Estatísticas básicas das avaliações.

#Sentenças	# Candidatos	#Avaliadores	#Avaliações
256	523	176	1.677

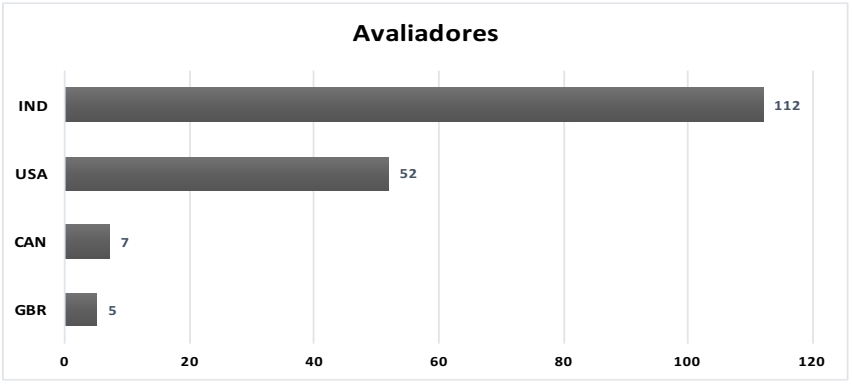


Figura 14 – Localização dos avaliadores do método proposto.

Quando a avaliação é concluída, a ferramenta figure-eight agrega os resultados dos três avaliadores com uma pontuação de confiança[7]. O índice de confiança descreve o nível

[7] https://success.figure-eight.com/hc/en-us/articles/201855939-How-to-Calculate-a-Confidence-Score

de concordância entre os avaliadores (ponderado pelas pontuações de confiança de cada avaliador) e indica a confiança na validade da resposta agregada para cada questionário. O resultado agregado é escolhido com base na resposta com a maior confiança.

Considerando essas sugestões, e para mostrar os resultados da avaliação intríseca na próxima seção, vale a pena relembrar a hipótese apontada neste capítulo: o resumo extrativo automático pode ser melhorado reduzindo-se as sentenças, permitindo a diminuição da redundância e mantendo a fluência e informatividade, ou até mesmo aumentando-se a informatividade com a possibilidade de inserção de novas palavras ou sentenças até o limite da taxa de compressão estabelecido.

5.2.2.1 Avaliação Intrínseca

A avaliação intrínseca visa inferir a qualidade da sentença, não especificamente um resumo, que foi reduzida usando a abordagem proposta na Seção 5.1. A tabela 34 mostra os resultados da avaliação para as 256 sentenças. Para a primeira pergunta, os avaliadores disseram que 249 sentenças preservam a mesma idéia da sentença original, isso equivale a 97% de precisão. Já para segunda pergunta foram 241 (95%) sentenças reduzidas que podem substituir as originais, além de preservar a mesma ideia. Analisando em detalhes, percebemos que cinco sentenças reduzidas preservaram a mesma ideia da original, porém segundo os avaliadores não poderiam substituir as originais. A tabela 35 mostra essas sentenças.

O método proposto conseguiu uma redução satistafória do conteúdo das sentenças, sem perder a qualidade. Para complementar os resultados da avaliação manual, a tabela 33 mostra em detalhes a média de palavras no corpus antes da compressão e depois da compressão. O método proposto conseguiu uma redução de aproximadamente 52% de redução do número de palavras. Na tarefa de sumarizacão, isso possibilidade a inserção de novas informação ao sumário gerado. Esses resultados (tabela 33) e os resultados da avaliação manual (tabela 34) apoiam a hipótese deste trabalho.

Tabela 33 – Resultados estatísticos da redução: média de palavras das sentenças originais e reduzidas, percentual de redução e desvio padrão entre parênteses

Sentenças originais	**Sentenças reduzidas**	**Redução (%)**
20,54 (5,99)	10,25 (3,23)	51,98 (15,20)

Tabela 34 – Avaliação humana do método de redução de sentenças.

	Sim	**Não**
A sentença simplificada preserva a mesma ideia da original?	249 (97%)	7 (3%)
A sentença simplificada pode substituir a original sem perder informações?	241 (95%)	13 (5%)

Tabela 35 – Sentenças avaliadas que não poderia substituir as sentenças originais.

Sentença original	**Sentença reduzida**
Iraqis braved the threat of bombs and other violence to vote Wednesday in parliamentary elections amid a massive security operation as the country slides deeper into sectarian strife.	Iraqis braved the threat of bombs and other violence.
The Parliamentary Standards Commissioner says there is “insufficient evidence” to merit an inquiry into the Conservative MP for the Wrekin, Mark Pritchard.	The Parliamentary Standards Commissioner says there is “insufficient evidence”.
Jonathan Martin has been placed on the reserve/non-football illness list on Saturday, ending the season for the embattled Miami Dolphins offensive lineman.	Jonathan Martin has been placed on the reserve/non-football illness list on Saturday.
China Wednesday donated a sea wall to Fiji to check coastal erosion.	China Wednesday donated a sea wall to Fiji.
Ken Norton who fought Muhammad Ali three times, has died at the age of seventy.	has died at the age of seventy.

A taxa de acerto é calculada a partir dos “sim” dos três avaliadores para uma dada sentença. Sendo que esses avaliadores são filtrados com base nos parâmetros apresentados na seção anterior. Portanto, o método proposto conseguiu resultados satisfatórios com acurácia geral de 95%. A tabela 37 mostra algumas sentenças reduzidas que preservam a mesma ideia das sentenças originais sem perder a informatividade. Sendo assim, essas sentenças podem substituir as originais e, consequentemente, possibilita inserir novas sentenças no sumário até o limite da taxa de compressão.

Tabela 37 – Exemplos de sentenças reduzidas com sucesso.

Sentença original	**Sentença reduzida**
Big Cinemas, a division of Reliance MediaWorks, has entered washroom advertising.	Big Cinemas has entered washroom advertising.
In an unusual case out of the United Kingdom, a British woman was jailed for trolling herself on Facebook.	A British woman was jailed for trolling herself on Facebook.
The Pendleton Round-Up, now 103 years old, plans to hire its first paid executive.	The Pendleton Round-Up plans to hire its first paid executive.
The Mercedes C-Class interior has been revealed, after some pictures appeared on a blog site.	The Mercedes C-Class interior has been revealed.
Kenyan police have repatriated more than 60 Ethiopia aliens to ease congestion at the local police cells.	Kenyan police have repatriated more than 60 Ethiopia aliens.
Musician Phil Everly, the younger half of The Everly Brothers, died on Jan. 3, 2014, in Burbank.	Musician Phil Everly died on Jan. 3, 2014, in Burbank.

5.3 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este capítulo apresentou a proposta de uma abordagem baseada em regras para redução de sentenças como etapa de pós-processamento de sumarização extrativa. A abordagem proposta tem como objetivo maximizar a informatividade do resumo, bem como evitar a redudância e manter a mesma qualidade textual (fluência, coesão e legibilidade) do sumário original.

Os resultados experimentais, em termos de avaliação manual e automática, demostram que a abordagem proposta é viável e apresentou resultados satisfatórios. Além disso, a avaliação humana apresentada nesta tese é cinco vezes maior que o trabalho de Filippova e Altun (2013). Elas avaliaram a qualidade da compressão automática para construção do corpus com apenas 50 sentenças. O autor desta tese fez uma avaliação manual seguindo uma rigososa metodologia (descrita na Seção 5.2.1) com 176 avaliadores qualificados para 256 sentenças selecionadas aleatoriamente de um corpus com 250.000 sentenças.

Analisando os resultados obtidos neste capítulo, é possível observar que estratégias que usam características léxicas são importantes para redução de sentenças. O autor desta tese acredita que sumarizadores automáticos de texto podem se beneficiar da abordagem proposta neste trabalho.

6 MEDIDAS DE QUALIDADE DE SUMÁRIOS

O capítulo 2 faz uma extensa avaliação quantitativa e qualitativa de diversos sistemas do estado da arte. Os capítulos 3, 4 e 5 propõem diferentes abordagens para melhoria na qualidade textual dos sumários extrativos como etapa de pós-processamento. Entretanto, as abordagens propostas são estáticas no sentido de que são aplicadas a único sumário de entrada. A revisão literária apresentada no capítulo 2 mostrou que a maioria dos trabalhos preocupam-se apenas com a informatividade dos sumários e poucos trabalhos usam abordagens dinâmicas para seleção de um sumário a partir de um conjunto de sumários gerados.

Alguns trabalhos da literatura (HONG; MARCUS; NENKOVA, 2015; WAN et al., 2015) mostram que a síntese de documentos com características diversas, tal como textos de notícias, usando uma única abordagem de sumarização extrativa possui uma significante limitação. Segundo Oliveira et al. (2017), isso acontece porque um único método de sumarização extrativa não consegue manter um alto nível de qualidade 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.

Oliveira et al. (2016a) avaliou e fez combinações de dezoito técnicas de sumarização e concluiu que nenhum dos métodos, combinações e sistemas analisados conseguiu manter um alto desempenho para todos os documentos ou grupos de documentos. Isso confirma os trabalhos anteriores de Ferreira et al. (2013), Ferreira et al. (2014) que 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).

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; DEWALIYA; GOPALANI, 2015), ordenamento (WAN et al., 2015) ou estimação da informatividade (HONG; MARCUS; NENKOVA, 2015) desses resumos candidatos. Essas abordagens 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. Outro fator não menos importante e explorado neste capítulo é a qualidade textual (fluência, coesão e legibilidade) dos resumos gerados que essas abordagens não levam em consideração no processo de combinação ou na estimação do melhor resumo.

No campo de estimar a informatividade de um resumo, Hong, Marcus e Nenkova (2015) e

Oliveira et al. (2017) demonstraram que adotar algoritmos de regressão apresentou melhores resultados do que adotar algoritmos de classificação de sentenças para a sumarização monodocumento ou 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 classificá-los ou combiná-los.

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. Oliveira et al. (2017) apresentou uma abordagem similiar, baseada em conceitos utilizando programação linear inteira e regressão para as tarefas de sumarização monodocumento e multidocumento de artigos de notícias. Ele apresentou resultados satisfatórios para 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.

Contudo, analisando os resultados obtidos por Hong, Marcus e Nenkova (2015), Wan et al. (2015), Oliveira et al. (2017), esses podem ser complementados com medidas para estimar a qualidade dos resumos gerados. Em muitos casos, os resumos podem ter um alto nível de informatividade, mas com baixo nível de compreensão, dificultando o entendimento pelo leitor.

Em uma linha de pesquisa na mesma direção à proposta neste capítulo, voltada para a avaliação de sistemas de SAT, Louis e Nenkova (2009a) e 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 existe uma boa correlação das características investigadas com as medidas de avaliação do PYRAMID e do ROUGE.

A avaliação automática de qualidade de sumários neste capítulo visa suprir as limitações supracitadas do método proposto por Hong, Marcus e Nenkova (2015) e Oliveira et al. (2017). Neste contexto, este capítulo busca construir: (i) uma abordagem para avaliação automática de sumários candidatos independente de resumos de referência; (ii) uma extensa avaliação automática do desempenho dos dez melhores sistemas de sumarização avaliados nesta tese; (iii) investigação e avaliação das principais medidas de informatividade e qualidade textual do estado arte; e (iv) uma extensa avaliação manual de 2.050 sumários seguindo uma criteriosa metodologia para avaliar a legibilidade, fluência, clareza referêncial e coesão. A arquitetura da solução proposta é composta por duas etapas principais: Geração dos resumos candidatos e Seleção do resumo mais informativo, descritas em detalhes na

próxima seção.

O restante deste capítulo está estruturado 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 automaticamente e manualmente as medidas e sistemas selecionados. Finalmente, a Seção 6.3 apresenta as considerações finais deste capítulo.

6.1 ABORDAGEM PROPOSTA

A metodologia de sumarização desenvolvida neste capítulo propõe uma abordagem indenpendente de sistemas de sumarização extrativa para avaliação automática de resumos. A escolha do sumário candidato é feita por um algoritmo de análise de qualidade textual que usa medidas que não necessitam de um sumário de referência(LOUIS; NENKOVA, 2013a). A figura 15 mostra uma visão geral da abordagem proposta, que consiste nas seguintes etapas: Pré-Processamento, Geração dos resumos candidatos e Seleção do resumo mais informativo e coesivo.

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.

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

6.1.1 Geração de resumos candidatos

Esta etapa é responsável pela geração de um conjunto de resumos candidatos a partir de um único documento (monodocumento) de entrada. Para isso, os 10 melhores sistemas avaliados nos Capítulos 2 e 4 são utilizados.

Dois critérios são utilizados para considerar os resumos de um sumarizador: (i) ser extrativo e monodocumento; e (ii) gerar sumários corretamente com base na taxa de compressão definida. Baseado nesses critérios, qualquer sistema de sumarização pode ser usada nessa etapa, o que corrabora para uma abordagem independente de sumarizador.

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

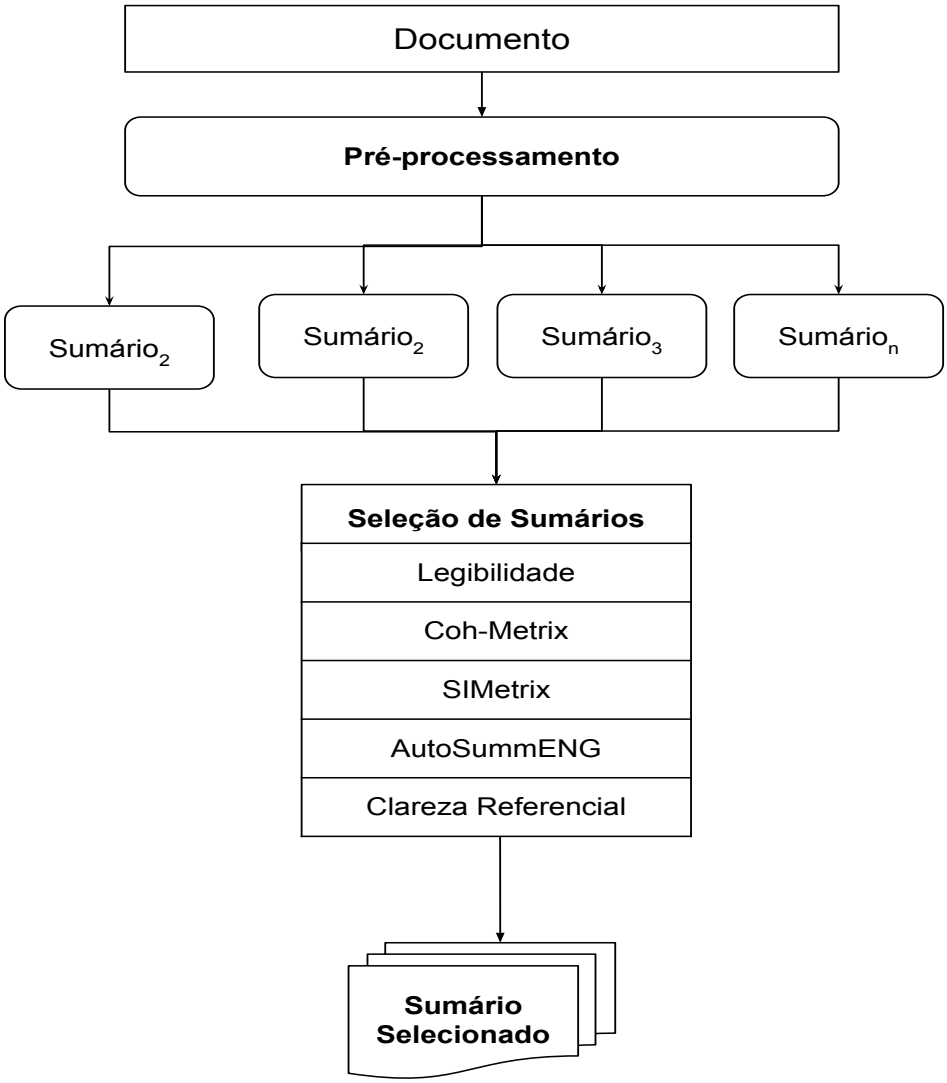


Figura 15 – Visão geral da abordagem proposta.

6.1.2 Seleção do resumo independente do resumo de Referência

Para seleção automática do resumo mais informativo e fluente, foram avaliadas diversas medidas tradicionais do estado da arte ou que foram propostas recentemente para estimar a qualidade textual do resumo (STEINBERGER, 2007; LOUIS; NENKOVA, 2008; LOUIS; NENKOVA, 2009b; LOUIS; NENKOVA, 2009a; LOUIS; NENKOVA, 2013a; MACKIE et al., 2014; MCNAMARA et al., 2014). Essas medidas foram utilizadas para avaliar e selecionar o sumário candidato no nível de informatividade, coesão, coerência de conteúdo e legibilidade.

As subseções a seguir descrevem as medidas selecionadas e que foram implementadas nesta tese:

6.1.2.1 Legibilidade (*Readability*)

Segundo (SCHRIVER, 1989), a definição do que pode ser considerado um texto bem escrito e legível depende fortemente do público-alvo. Por exemplo, um artigo científico bem escrito

não será muito legível para uma pessoal leiga, assim como um romance poderá ser pouco apreciado por um aluno da terceira série (PITLER; NENKOVA, 2008).

Há diversos trabalhos na literatura que medem a qualidade de um texto com base nas medidas de legibilidade (PITLER; NENKOVA, 2008; LOUIS, 2012; NANDHINI; BALASUNDARAM, 2013; LOUIS; NENKOVA, 2013b). Essas medidas possibilitam avaliar a legibilidade de um texto. Para cada medida há indicadores para avaliar a capacidade de leitura, geralmente por contagem de sílabas, palavras e sentenças. Percebe-se que em grande parte o vocabulário determina o nível de legibilidade de um texto.

Nesta tese, foram pesquisadas as principais medidas de legibilidade da literatura. Elas foram divididas em dois grupos: facilidade de leitura e níveis de classificação.

Facilidade de leitura (*Reading Ease*)

A medida Flesch-Kincaid Reading Ease (FRES) (KINCAID et al., 1975) foi selecionada para indicar o nível de facilidade para leitura de um texto. Uma elevada pontuação indica o quão maior a legibilidade. As pontuações variam geralmente entre 0 e 100. Um texto que tem uma pontuação entre 60 e 80 é considerado de fácil leitura.

A fórmula 6.1 mostra os detalhes de como são calculadas as pontuações de *Flesch-Kincaid Reading Ease*.

$$206.835 - 1.015 * \left(\frac{palavras}{sentenças}\right) - 84.6 * \left(\frac{sílabas}{palavras}\right) \tag{6.1}$$

Níveis de Ensino (*Grade Levels*)

Um nível de ensino (baseado no sistema educacional dos Estados Unidos) equivale ao número de anos de estudo que uma pessoa tem, por exemplo, uma pontuação entre 10 e 12 equivale ao nível de leitura de uma pessoa que concluiu o ensino médio. Textos voltados para o público em geral (por exemplo, notícias) devem ter um nível de ensino próximo de 8.

Nesse contexto, foram selecionadas algumas medidas tradicionais do estado da arte com intuito de avaliar o nível de leitura dos sumários. Espera-se que um bom sumário tenha um nível de leitura próximo ao seu documento original. A seguir são descritas brevemente essas medidas:

- Nível de Flesch-Kincaid (*Flesch-Kincaid Grade Level*) (F-K GL) (KINCAID et al., 1975): Esta medida de legibilidade é usada extensivamente no campo da educação. F-K GL apresenta uma pontuação para definir o nível de escolaridade baseado no sistema educacional dos Estados Unidos, facilitando para professores, pais, bibliotecários e outros avaliarem o nível de legibilidade de vários livros e textos. Também pode significar o número de anos no sistema educacional necessários para entender um dado texto, esse valor é relevante quando a fórmula resulta em um número maior que

10 (correspondente ao último ano da educação primária). O nível de escolaridade é calculado com a seguinte fórmula:

$$0.39 * \left(\frac{palavras}{sentenças}\right) + 11.8 * \left(\frac{sílabas}{palavras}\right) - 15.59 \tag{6.2}$$

- Índice de Gunning Fog (*Gunning Fog Index*) (GFI) (GUNNING, 1952): Em linguística, o GFI é um teste de legibilidade para a escrita em inglês. O índice estima os anos de educação formal que uma pessoa precisa para entender o texto na primeira leitura. Por exemplo, um GFI com pontuação 12 requer o nível de leitura de um aluno do ensino médio dos Estados Unidos (com cerca de 18 anos de idade). O GFI é comumente usado para confirmar que o texto pode ser lido facilmente pelo público-alvo. A fórmula é definida como:

$$0.4 * \left(\left(\frac{palavras}{sentenças}\right) + 100 * \left(\frac{palavrascomplexas}{palavras}\right)\right) \tag{6.3}$$

- Índice de Coleman-Liau (*Coleman-Liau Index*) (CLI) (COLEMAN; LIAU, 1975): O CLI é um teste de legibilidade projetado por Meri Coleman e Liau para avaliar a compreensibilidade de um texto. Assim como as demais medidas de legibilidade, o CLI busca relacionar a facilidade de leitura de um texto ao nível educacional do leitor. No entanto, Coleman-Liau se baseiam em caracteres em vez de sílabas por palavra, conforme fórmula abaixo:

$$\left(5.89 * \frac{caracteres}{palavras}\right) - 0.3 * \left(\frac{sentenças}{palavras}\right) - 15.8 \tag{6.4}$$

- Índice SMOG (*SMOG Index*) (SI) (MCLAUGHLIN, 1969): É uma medida de legibilidade que estima os anos de escolaridade necessários para entender um determinado texto. A fórmula para calcular o SI foi desenvolvida por G. Harry McLaughlin como um substituto mais preciso e mais facilmente calculado para o GFI:

$$1.043 * \sqrt{palavrascomplexas * \left(\frac{30}{sentenças}\right)} + 3.1291 \tag{6.5}$$

- Índice de legibilidade automatizada (*Automated Readability Index*) (ARI) (KINCAID et al., 1975): O ARI é um teste de legibilidade para textos em inglês, projetado para avaliar a compreensibilidade de um texto. A fórmula para calcular o ARI é dada abaixo:

$$4.71 * \left(\frac{caracteres}{palavras}\right) + 0.5 * \left(\frac{palavras}{sentenças}\right) - 21.43 \tag{6.6}$$

- Grau de Spache (*Spache score*) (SS) (SPACHE, 1953): É uma medida de legibilidade criada por George Spache para dar suporte a escrita de textos em inglês. Essa medida

tem melhor desempenho em textos voltados para crianças até a quarta série do ensino primário. A fórmula é definida a abaixo:

$$(0.121 * AvgTamanhoSentença) + (0.082 * \%PalavrasIncomuns) + 0.659 \quad (6.7)$$

- *Dale–Chall score* (D-C S) (DALE; CHALL, 1948): É uma medida de legibilidade que fornece um indicador numérico da dificuldade de compreensão que os leitores encontram ao ler um dado texto. A fórmula para calcular a pontuação D-C S é dada abaixo:

$$0.1579 * \left(\frac{palavrascomplexas}{palavras} * 100\right) + 0.0496 * \left(\frac{palavras}{sentenças}\right) \quad (6.8)$$

- *Average Grade Level*: média geral das medidas descritas acima.

$$\frac{(FKGL + GFI + CLI + SI + ARI + SS + DCS)}{7} \quad (6.9)$$

6.1.2.2 Coh-Metrix

Coh-Metrix[1] é uma sofisticada ferramenta de avaliação automática do texto e discurso (MCNAMARA et al., 2014). É possível dimensionar a dificuldade ou a facilidade para compreensão de um texto por meio de 108 medidas[2] para representações linguísticas e discursivas. Além disso, ela pode ser utilizada de muitas maneiras diferentes para investigar a coesão e coerência da representação mental de um texto. Graesser, McNamara e Louwerse (2003) definem coesão como características explícitas do texto que desempenham algum papel em ajudar o leitor a conectar mentalmente as ideias no texto.

Nos estudos experimentais de Dowell, Graesser e Cai (2015a) e Graesser, McNamara e Kulikowich (2011), demonstraram que apenas oito medidas[3] são suficientes para representar 67% da variância em 37.520 textos do corpus da TASA (*Touchstone Applied Science Association*), que representa o que estudantes típicos do ensino médio leram (romances, artigos de jornais, etc.) ao longo da vida (DOWELL; GRAESSER; CAI, 2015b).

Esses indicadores estão notavelmente alinhados com a estrutura teórica multinível por Graesser e McNamara (2011). Além disso, esses indicadores são calculados na forma de *z-scores* e percentuais. *Z-score* é uma pontuação padrão que indica quantos desvios-padrão uma observação ou dado está acima ou abaixo da média, onde a média é definida como 0. Já o percentual varia de 0 a 100%, com pontuações mais altas significando que o texto provavelmente será mais fácil de ler do que outros textos.

Nesta tese, foram efetuadas as avaliações experimentais para as oito medidas e a pontuação de formalidade sugeridas por (DOWELL; GRAESSER; CAI, 2015a):

[1] http://cohmetrix.com/
[2] http://tool.cohmetrix.com/
[3] http://tea.cohmetrix.com/

1. Narratividade - *Narrativity* (PCNARz, PCNARp). O texto narrativo conta uma história, com personagens, eventos, lugares e coisas que são familiares para o leitor. A narrativa é intimamente afiliada à conversação oral e cotidiana. Este componente robusto é altamente afiliado com o domínio da palavra, conhecimento do mundo e linguagem oral. Textos não narrativos sobre tópicos menos conhecidos estão no extremo oposto.

2. Simplicidade Sintática - *Syntactic Simplicity or Syntactic ease* (PCSYNz, PCSYNp). Esse componente reflete o grau em que as sentenças no texto contêm poucas palavras e usam estruturas sintáticas mais simples e conhecidas, que são mais fáceis de processar e entender. No extremo oposto estão os textos que contêm frases com mais palavras e que usam estruturas sintáticas complexas e desconhecidas.

3. Concretude da Palavra ou Concretização de palavras - *Word Concretenes* (PCCNCz, PCCNCp). Textos que contêm palavras de conteúdo que são concretas e significativas e evocam imagens mentais são mais fáceis de processar e entender. Palavras abstratas representam conceitos que são difíceis de representar visualmente. Textos que contêm palavras mais abstratas são mais difíceis de entender.

4. Coesão Referencial - *Referential Cohesion* (PCREFz, PCREFp). Um texto com alta coesão referencial contém palavras e ideias que se sobrepõem às sentenças e a todo o texto, formando linhas explícitas que conectam o texto ao leitor. Normalmente, o texto de baixa coesão é mais difícil de processar porque há menos conexões que unem as ideias para o leitor. Portanto, esse componente reflete o quanto as palavras e ideias explícitas no texto estão conectadas umas às outras à medida que o texto se desdobra.

5. Coesão Profunda - *Deep Cohesion* (PCDCz,PCDCp). Essa dimensão reflete o grau em que o texto contém conectivos causais, intencionais e temporais. Esses conectivos ajudam o leitor a formar uma compreensão mais profunda e coerente dos eventos, processos e ações causais no texto. Quando um texto contém muitos relacionamentos, mas não contém esses conectivos, o leitor deve inferir as relações entre as ideias. Se o texto é rico em coesão profunda, então essas relações e a coesão global são mais explícitas.

6. Coesão verbal (*Verb Cohesion*) (PCVERBz, PCVERBp). Esse componente reflete o grau em que há verbos sobrepostos no texto. Quando há verbos repetidos, o texto provavelmente inclui uma estrutura de eventos mais coerente que facilitará e melhorará o entendimento do modelo de situação. A pontuação desse componente é provavelmente mais relevante para textos destinados a leitores mais jovens e para textos narrativos (McNamara, Graesser, & Louwerse, 2012).

7. Conectividade (*Connectivity*) (PCCONNz, PCCONNp). Esse componente reflete o grau em que o texto contém conectivos adversativos, aditivos e comparativos explícitos para expressar as relações no texto. Esse componente reflete o número de relações lógicas no texto que são explicitamente transmitidas. Essa pontuação provavelmente está relacionada à compreensão mais profunda do leitor sobre as relações no texto.

8. Temporalidade (*Temporality*) (PCTEMPz, PCTEMPp). Textos que contêm mais sugestões sobre temporalidade e que têm temporalidade mais consistente (isto é, tempo, aspecto) são mais fáceis de processar e entender. Além disso, a coesão temporal contribui para o entendimento do nível de compreensão do leitor sobre os eventos no texto.

Além das oito medidas descritas acima, Graesser et al. (2014) propuseram uma medida conhecida como formalidade (BIBER, 1988; CRYSTAL, 1987; LABOV, 1972; HEYLIGHEN; DEWAELE, 2002; HEYLIGHEN, 1999) para dimensionar dificuldade de compreensão de um texto. O discurso formal tem um estilo técnico empolado que é apropriado quando há necessidade de ser preciso, coerente, articulado e convincente para um público instruído. A linguagem formal é geralmente encontrada na mídia impressa (como artigos acadêmicos, enciclopédias, jornais e textos educacionais) ou oratória pré-planejada (por exemplo, a comunicação de precisão crítica usada durante o controle de tráfego aéreo, discursos políticos e notícias jornalísticas faladas). Esses gêneros de discurso são cuidadosamente elaborados e editados, em vez de serem preparados extemporaneamente - como no caso do discurso.

Portanto, calculou-se uma pontuação de formalidade baseada em cinco medidas das oito dimensões definidas anteriormente, conforme a Equação 6.10.

$$Formalidade = \frac{PCNARz + PCREFz + PCSYNz + PCCNCz + PCDCz}{5} \quad (6.10)$$

6.1.2.3 *SIMetrix: Summary Input similarity Metrics*

SIMetrix é uma ferramenta desenvolvida por Louis e Nenkova (2013a) para avaliação de sumários candidatos sem a necessidade de um sumário de referência. Basicamente, ela compara os resumos gerados com os documentos de origem.

Segundo Louis e Nenkova (2013a), uma vez que os resumos são substitutos dos documentos de origem, uma alta similaridade entres eles seria um indicativo de sumários com boa qualidade.

A SIMetrix foi testada por meio de correlações de avaliadores humanos com medida ROUGE (LOUIS; NENKOVA, 2008; LOUIS; NENKOVA, 2009b; LOUIS; NENKOVA, 2009a). Os autores obtiveram resultados satisfatórios, no melhor resultado alçaram 93% de correlação

com humanos e está em paridade com medidas de avaliação automática que usam sumários de referência (por exemplo, ROUGE).

Mackie et al. (2014) avaliou o desempenho do SIMetrix e do ROUGE para sumarização de *Microblog*. Os autores usaram duas medidas do SIMetrix: *Jensen-Shannon Divergence* e *Fraction of Topic Words*. Eles demonstraram que a medida *Fraction of Topic Words* superou o ROUGE-1 em relação a concordância com avaliadores humanos.

As medidas selecionadas para avaliação neste trabalho são classificadas em quatro grupos (LOUIS; NENKOVA, 2013a):

Similaridade de Distribuição (*Distribution Similarity*)

Neste grupo, Louis e Nenkova (2013a) avaliaram três medidas de similaridade entre duas distribuições de probabilidade: *Kullback Leibler Divergence* (KLD), *Jensen Shannon Divergence* (JSD) e Similaridade do Cosseno (cosineOverlap). Para caracterizar um bom resumo, espera-se baixa divergência entre as distribuições de probabilidade das palavras do documento de entrada e do resumo, e alta similaridade entre eles. Estas foram utilizadas para avaliação de resumo em outros trabalhos na literatura (DONAWAY; DRUMMEY; MATHER, 2000; LIN et al., 2006; STEINBERGER et al., 2007), embora em um contexto diferente. Esses autores avaliaram a similaridade entre sumários candidatos e os produzidos por humanos.

Semelhança entre Resumos (*Summary Likelihood*)

As medidas baseadas na semelhança entre resumos *summary likelihood* são obtidas através de um modelo gerador da probabilidade de palavras no documento de entrada. Esse modelo é usado para calcular a probabilidade da semelhança entre os resumos. Sob este modelo generativo, a probabilidade do conteúdo de um resumo pode ser calculada usando diferentes métodos, espera-se maior probabilidade para resumos com melhor qualidade.

Nesta tese, foram implementados dois modelos propostos por Louis e Nenkova (2013a):

- *Unigram summary probability:* Presume-se que a probabilidade de uma palavra no texto de referência represente a probabilidade da palavra aparecer em um resumo. Essa medida é conhecida como probabilidade de unigrama do resumo.; e
- *Multinomial summary probability:* semelhante a probabilidade de unigrama, porém usa-se uma distribuição multinomial (FORBES et al., 2011).

Medidas baseadas em palavras descritivas (*Topic's Words in the Summary*)

Uma sentença é pontuada conforme o número de palavras descritivas que apresenta. As palavras descritivas também são conhecidas como *topic words*, *topic signatures* ou *signature*

terms[4]). Sistemas que otimizam diretamente o número de *topic words* têm se saído muito bem nas avaliações (CONROY; SCHLESINGER; O'LEARY, 2006; RIBALDO; CARDOSO; PARDO, 2016). Portanto, o número de palavras descritivas presente em um resumo pode ser um bom indicativo da qualidade do seu conteúdo (LOUIS; NENKOVA, 2013a).

Foram implementadas duas medidas que quantificam a presença de *topic words*:

1. *Fraction of Topic Words* (FoTW) mede o quociente de palavras de tópicos derivadas dos documentos de entrada que também são encontrados no resumo. Essa medida foi originalmente proposta como característica (*feature*) de sumarização (LIN; HOVY, 2000).

2. Percentual de *topic signatures* do documento de entrada que também aparecem no resumo.

Ambas as medidas têm pontuações mais altas para sumários que contêm muitas *topic words*. A primeira medida é guiada apenas por qualquer *topic words* (palavra, radical, bigrama ou trigrama) e a segunda mede a diversidade de *topic words* utilizadas no sumário.

6.1.2.4 Método AutoSummENG

O método AutoSummENG, proposto por Giannakopoulos et al. (2008), baseia-se no conceito de usar informações textuais extraídas estatisticamente para medir a similaridade entre resumos gerados e os resumos de referência.

Este método é baseado em grafos de n-gramas e leva em conta uma ocorrência simultânea de uma palavra ou uma sequência de caracteres. Na abordagem baseada em palavras, é sempre desejável utilizar a forma base (*lemma*) da palavra, isto é, converter todas as formas diferentes de uma palavra em sua forma base (por exemplo, converter *goes* para *go*). Isso requer lematizadores que não estão disponíveis para muitos idiomas, mas apenas para alguns idiomas, como inglês, alemão, etc. Para manter essa abordagem independente da linguagem, os autores simplesmente usam n-gramas de caracteres (sequências de caracteres) no cálculo da co-ocorrência. Este método mostrou ter maior correlação com julgamentos humanos do que o ROUGE (GIANNAKOPOULOS; KARKALETSIS, 2011).

No contexto desta tese, o método AutoSummENG foi avaliado para medir a similaridade entre os resumos candidatos e o texto original. Além disso, definiram-se n-gramas de tamanho máximo igual a três para representação dos grafos. A partir dessa representação, as seguintes medidas foram usadas para calcular a similaridade entre o resumo e a notícia (GIANNAKOPOULOS et al., 2008; GIANNAKOPOULOS; VOUROS; KARKALETSIS, 2010):

[4] É como são chamadas as palavras descritivas em alguns trabalhos da área e que fazem uso de função de verossimilhança

1. *Value Similarity* (VS): calcula a similaridade entre os grafos do resumo (G_1) e do texto original (G_2), onde VS igual a 100% indica correspondência perfeita entre os grafos comparados;

2. *Size Similarity* (SS): razão entre o grafo com menor número de arestas e o grafo com maior número de arestas;

3. Containment Similarity (CS): percentual de arestas em comum entre os grafos (co-ocorrência); e

4. *Overall Similarity* (OS): média ponderada das três medidas acima, que corresponde à similaridade geral entre o resumo e o texto original.

6.1.2.5 Expressões Anafóricas Quebradas ou Clareza Referencial (*Reference Clarity*)

A última medida implementada na presente tese foi baseada na metodologia proposta por Smith, Henrik e Arne (2012). Essa medida foi amplamente utilizada no Capítulo 3. Ela contabiliza o número de expressões anafóricas quebradas no resumo, ou seja, pronomes e substantivos devem ser claramente referidos no resumo. Para essa medida, quanto maior o número de expressões anafóricas quebradas, menor é a qualidade de um resumo.

6.2 EXPERIMENTOS

Esta seção apresenta e discute os experimentos realizados para avaliar diferentes aspectos da avaliação automática de sumários. Os experimentos foram conduzidos buscando avaliar manualmente o desempenho das medidas automáticas para seleção de sumários indenpendente de um sumário de referência.

6.2.1 Configurações dos experimentos

Esta seção descreve as configurações dos experimentos que compreendem a ferramenta de avaliação humana e a metodologia de avaliação utilizada.

Ferramenta de avaliação. De maneira similar à explicada no capítulo anterior, foi aqui adotada a ferramenta *figure-eight*[5] para avaliação humana. Ela usa o conceito de *Human-in-the-loop* (HITL) para criar e validar modelos de aprendizagem de máquina usando inteligência humana e de máquina. Em uma abordagem tradicional de HITL, as pessoas estão envolvidas em um círculo virtuoso no qual elas treinam, ajustam e testam um algoritmo específico.

Figure-eight permite criar grandes quantidades de dados de treinamento altamente precisos para um dado caso de uso, ajustar modelos com percepção humana e testá-los para garantir que as decisões sejam precisas e acionáveis.

[5] https://www.figure-eight.com/

Metodologia de avaliação. Foram selecionados de forma aleatória 205 documentos do corpus CNN e os 10 melhores sumarizadores automáticos avaliados nos capítuos anteriores desta tese. Os resumos gerados utilizaram a taxa de compressão de 10% do número de sentenças do documento de entrada. Foram gerados um total de 2.050 sumários para serem avaliados manualmente. Além disso, para cada sumário foi analisado automáticamente o nível de informatividade e fluência (coesão e legibilidade) usando as medidas automáticas descritas na Seção 6.1.2. Ao final, analisou-se a correlação entre as medidas automáticas e as avaliações manuais.

Durante a etapa de avaliação manual aplicaram-se questionários para avaliar os resumos em relação a fluência e a informatividade. Para garantir a qualidade do processo de avaliação configuraram-se critérios na ferramenta figure-eight semelhantes aos já descritos no experimento do capítulo anterior:

1. permitir apenas avaliadores de 9 países (Estados Unidos, Canadá, Austrália, Reino Unido, Bahamas, Barbados, Índia, Jamaica e Nova Zelândia), cuja língua oficial é o inglês;

2. os avaliadores devem possuir uma acurácia mínima de 70%. Essa porcentagem é a precisão mínima que um avaliador deve manter durante a avaliação para continuar avaliando. Se o avaliador ficar abaixo dessa precisão a qualquer momento, ele ou ela será removido do trabalho e todas as suas respostas serão desconsideradas ou não confiáveis;

   Para calcular essa porcentagem, foram utilizadas algumas perguntas de teste com casos positivos e negativos em cada questionário. Sendo que para cada avaliação válida, o avaliador deveria responder no mínimo uma pergunta de teste;

3. um número de 3 avaliadores para cada questionário. No final é calculado um percentual de confiança (descrito na próxima seção);

4. o tempo mínimo de avaliação por questionário foi de 1 minuto;

5. cada avaliador só pode responder no máximo 10 questionários. Sendo 1 documento com 10 resumos por questionário; e

6. todos os avaliadores devem ser nível 2 na ferramenta. Esse nível indica maior qualidade, ou seja, consiste num grupo menor de colaboradores mais experientes e com maior precisão.

   A figure-eight possue três níveis de avaliadores, sendo que o nível 3 possui avaliadores mais qualificados na ferramenta e consequentemente as avaliações com maior preço.

6.2.2 Resultados das medidas automáticas para seleção de resumos

O objetivo deste experimento é avaliar o desempenho de diferentes sistemas de sumarização segundo as medidas automáticas para avaliação de sumários. Essas medidas têm como objetivo indicar os sumários com maior nível de informativdade e/ou fluência.

A tabela 39 apresenta os resultados para as medidas de informatividade (cobertura, precissão e f-measure) do Rouge 1 (R-1) e Rouge 2 (R-2). Segundo a medidade de f-measure, o sistemas que apresentaram os melhores resultados foram PLI + GE, RL + PLI e MEAD, respectivamente. Essas medidas, apesar de serem muito utilizadas no estado da arte, não há uma noção de qualidade textual do resumo gerado. Elas foram utilizadas neste capítulo como referência comparativa com as demais medidas selecionadas e a avaliação manual.

Tabela 39 – Desempenho (%) geral dos sumários gerados pelos sistemas selecionados com base nas medidas do R-1 e R-2. Os melhores resultados são destacados em negrito.

Sistema	**ROUGE-1**			**ROUGE-2**		
	R	**P**	**F1**	**R**	**P**	**F1**
ALMUS	45,70	50,12	47,81	30,53	35,18	32,69
AutoS	48,42	49,50	48,96	31,64	34,39	32,96
Aylien	51,95	45,60	48,57	33,82	31,35	32,54
C4J	45,49	50,39	47,82	30,71	35,23	32,81
HP-FS	50,83	44,83	47,64	33,57	31,54	32,53
MEAD	50,13	49,43	**49,78**	34,67	36,70	**35,66**
PLI + GE	56,93	44,67	**50,06**	39,84	32,49	**35,79**
RL + PLI	57,06	43,57	**49,41**	38,68	31,40	**34,66**
SUMMA	52,68	46,15	49,20	35,52	33,43	34,44
TextRank	49,02	46,03	47,48	32,28	31,39	31,83

Na tabela 40 são apresentados os resultados das medidas de legibilidade nos sumários gerados. Foram selecionadas nove medidas clássicas de legibilidade do estado da arte, e para cada medida foram selecionados os sistemas que produziram o maior número de sumários legíveis. Para determinar os três sistemas com maior nível de legibilidade para cada medida, calculou-se a legibilidade do corpus e selecionaram-se os sistemas que tiveram pontuação de legibilidade mais próxima do corpus.

Segundo as medidas de legibilidade, o sistema TextRank atingiu o melhor desempenho geral, apresentou melhores resultados para 8 das 10 medidas avaliadas, seguido dos sistemas AutoS e C4J com 7 medidas cada. Observando cada medida individualmente, percebe-se que as medidas D-C, Spache, SI, G-F I, F-K GL e A GL tiveram o mesmo comportamento, indicando os mesmos sistemas com melhores avaliações gerais.

Tabela 40 – Desempenho (%) geral dos sumários gerados pelos sistemas selecionados com base nas medidas de legibilidade. Os melhores resultados são destacados em negrito.

Sistema	**D-C**	**Spache**	**ARI**	**S I**	**C-L I**	**G-F I**	**F-K GL**	**FRE**	**A GL**
Corpus	10,34	6,93	11,21	13,35	11,96	14,58	11,25	51,53	11,38
ALMUS	14,46	16,51	50,46	25,67	12,64	46,09	41,89	30,17	29,67
AutoS	**11,10**	**8,21**	**16,35**	**15,64**	12,93	**18,59**	**15,17**	38,99	**13,99**
Aylien	11,40	8,60	17,21	16,15	12,64	19,54	16,06	36,55	14,51
C4J	**10,86**	**7,78**	14,18	**14,51**	12,23	**16,90**	**13,59**	**44,60**	**12,86**
HP-FS	11,26	8,55	**16,87**	15,83	12,32	19,14	15,75	38,42	14,25
MEAD	11,15	8,45	16,74	15,69	**12,12**	18,99	15,61	**39,52**	14,11
PLI + GE	11,26	8,65	17,56	16,11	**12,18**	19,74	16,27	37,53	14,54
RL + PLI	11,27	8,80	18,26	16,30	**12,10**	20,28	16,78	36,69	14,83
SUMMA	11,43	8,71	17,76	16,26	12,40	19,91	16,41	36,69	14,70
TextRank	**11,08**	**8,20**	**15,86**	**15,21**	12,40	**18,18**	**14,86**	**40,98**	**13,69**

Assim como as medidas de legibilidade, os resultados das medidas de qualidade textual da ferramenta Co-Metrix (veja tabela 41) foram analisados seguindo a hipótese de que quanto mais próximo forem os resultados da avaliação do sumário do documento de entrada, melhor será sua qualidade.

Conforme apresentados na tabela 41, o sistema C4J obteve o melhor desempenho geral, com melhores resultados para 7 das 9 medidas avaliadas. Em sequência o sistema Almus foi avaliado com melhor desempenho para 4 medidas, e os sistemas AutoS e HP-FS para 3 medidas.

Analisando individualmente a medida de formalidade (F) do Co-Metrix, o sistema Almus atingiu o melhor desempenho, seguido dos sistemas C4J e PLI + GE.

Tabela 41 – Desempenho (%) geral dos sumários gerados pelos sistemas selecionados com base nas medidas do Co-Metrix. Os melhores resultados são destacados em negrito.

Sistema	PCN	PCS	PCC	PCR	PCD	PCV	PCCO	PCT	F
CORPUS	18,93	21,18	83,96	85,60	49,37	88,08	12,68	33,42	0,44
ALMUS	14,35	**5,35**	**93,84**	91,55	34,19	**88,43**	27,64	40,75	**0,55**
AutoS	**18,62**	2,70	95,10	95,44	**39,42**	78,39	**24,44**	34,83	0,74
Aylien	**16,47**	2,44	96,84	93,44	32,20	85,46	31,64	**32,13**	0,69
C4J	16,40	**7,40**	**90,13**	**87,21**	**38,15**	**90,10**	28,03	39,58	**0,51**
HP-FS	14,32	**3,51**	95,47	90,43	31,75	**88,79**	29,17	**33,63**	0,57
MEAD	23,25	2,48	**94,29**	**87,94**	37,62	77,80	26,40	36,43	0,62
PLI + GE	15,12	2,48	96,62	89,62	34,08	90,85	26,66	**34,71**	**0,54**
RL + PLI	16,24	2,47	96,08	**89,41**	36,38	90,31	27,03	34,97	0,58
SUMMA	**19,74**	0,60	96,59	96,95	37,01	85,17	**25,35**	42,49	0,81
TextRank	22,92	3,26	94,87	97,60	**38,05**	83,07	**25,93**	41,54	0,81

As medidas da ferramenta SIMetrix foram divididas em duas categorias: medidas de divergência (tabela 42) e medidas de similaridade (Tabela 43). As medidas de divergência têm como objetivo avaliar a menor divergência entre o sumário e seu documento original, já as medidas de similaridade buscam o contrário, isto é, maior similaridade indica resumos com maior informatividade. A tabela 42 apresenta os resultados das medidas de divergência. Os sistemas Autos, MEAD e SUMMA atingiram os melhores resultados gerais para essas medidas. Já para os resultados de similaridade apresentados na tabela 43, o sistema atingiu o melhor resultado geral, seguido dos sistemas AutoS e SUMMA.

Tabela 42 – Desempenho geral (média) dos sumários gerados pelos sistemas selecionados com base nas medidas do SIMetrix. Essas medidas identificam como melhor sumário o que possui menor diferença com o documento de entrada. Os melhores resultados são destacados em negrito.

Sistema	KLIptSumm	KLSummIpt	unsmoothedJSD	smoothedJSD
ALMUS	2,21	1,19	0,34	0,30
AutoS	**2,14**	**1,06**	**0,30**	**0,27**
Aylien	2,19	1,12	0,33	0,29
C4J	2,20	1,22	0,35	0,31
HP	2,23	1,16	0,34	0,30
MEAD	**2,17**	**1,06**	**0,31**	**0,28**
PLI + GE	2,22	1,14	0,33	0,30
RL + PLI	**2,17**	1,10	0,32	0,29
SUMMA	**2,15**	**1,00**	**0,30**	**0,27**
TextRank	**2,17**	1,09	0,32	0,29

Tabela 43 – Desempenho geral (média) dos sumários gerados pelos sistemas selecionados com base nas medidas do SIMetrix. Essas medidas identificam como melhor sumário o que possui maior similaridade com o documento de entrada. Os melhores resultados são destacados em negrito.

Sistema	cosine	pTpTokens	fTpWords	tpWordO	ugProb	mnProb
AutoS	**0,72**	**0,30**	**0,68**	**0,67**	-156,03	-56,56
Aylien	0,69	**0,29**	0,64	0,64	-148,81	-56,20
C4J	0,65	0,25	0,61	0,58	**-132,10**	**-53,64**
HP	0,65	0,25	0,60	0,57	-151,04	-57,56
PLI + GE	0,65	0,24	0,61	0,56	-155,76	-59,24
RL + PLI	0,66	0,24	0,63	0,56	-163,16	-60,50
SUMMA	**0,74**	0,27	**0,69**	**0,67**	-174,94	-63,18
MEAD	0,69	0,26	0,65	0,61	-167,43	-61,43
ALMUS	0,66	0,27	0,61	0,60	**-124,54**	**-50,95**
TextRank	**0,72**	**0,32**	**0,66**	**0,68**	**-144,81**	**-55,03**

Analisando individualmente as medidas da ferramenta AutoSummENG apresentadas na tabela 44, o sistema SUMMA atingiu o melhor resultado para a medida OS, para as medidas VS e SS foi o sistema AutoS, e para medida CS o sistema RL+ PLI atingiu o melhor resultado. De maneira geral, os AutoS, SUMMA e MEAD foram indicados com melhor desempenho por um maior número de medidas.

Tabela 44 – Desempenho (%) geral dos sumários gerados pelos sistemas selecionados com base nas medidas do AutoSummENG. Os melhores resultados são destacados em negrito.

Sistema	OS	VS	CS	SS
ALMUS	**4,60**	16,77	97,81	21,45
AutoS	**5,51**	**19,70**	96,64	**25,28**
Aylien	2,90	14,51	99,42	19,22
C4J	2,79	14,02	99,39	18,49
HP-FS	3,08	14,95	**99,52**	19,50
MEAD	4,32	**17,78**	98,90	**22,95**
PLI + GE	3,44	15,88	**99,59**	20,64
RL + PLI	3,74	16,60	**99,60**	21,59
SUMMA	**4,72**	**18,97**	99,25	**24,19**
TextRank	3,17	15,17	99,23	20,03

A última medida avaliada foi a de clareza referencial, cujo objetivo é contar os sumários que possuem coreferências quebradas, por exemplo, um pronome não conectato ao seu substântivo. A tabela 45 apresenta os sistemas e o número de sumários com problemas de

clareza referêncial. Os sistemas C4J, HP-FS e ALMUS antigiram os melhores resultados com um menor número de sumários com correferências quebradas, respectivamente.

Tabela 45 – Desempenho (%) geral dos sumários gerados pelos sistemas selecionados com base na medida de clareza referencial. Os melhores resultados são destacados em negrito.

Sistema	**#Clareza Referencial**
C4J	**80**
HP-FS	**83**
ALMUS	**89**
PLI + GE	92
Aylien	101
AutoS	113
MEAD	113
RL + PLI	114
SUMMA	130
TextRank	131

6.2.3 Resultados das avaliações manuais

O experimento aqui descrito avalia manualmente o desempenho dos sistemas de sumarização extrativa selecionados, seguindo os critérios de informatividade e qualidade textual definidos no próprio questionário de avaliação. Para isso, adotou-se a avaliação intríseca baseada em questionários. Um exemplo de um questionário com as instruções no cabeçalho, o *link* para a notícia original e os sumários candidatos para avaliação pode ser visto abaixo.

Evaluation Of Automatic Text Summarization

Instructions: This survey consists in assessing summaries automatically produced; each summary summarizes the news.

Read the news and summaries below and select the best summaries, taking into account the following quality criteria:

The summary in which it considers easier to read and understand The summary in which it considers better organized The summary that has the highest quality (cohesion and coherence) The summary that best represents the original text The summary that contains more relevant information The summary that contains the least amount of repeated information The summary that contains the fewest pronouns without being connected to an entity (e.g., noun) And you should take into account the five aspects of linguistic quality described below:

1. **Grammaticality:** The summary should have no datelines, system-internal formatting, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.

2. **Non-redundancy:** There should be no unnecessary repetition in summary. Unnecessary repetition might take the form of whole sentences that are repeated, or repeated facts, or the repeated use of a noun or noun phrase (e.g., "Bill Clin- ton") when a pronoun ("he") would suffice.

3. **Referential clarity:** It should be easy to identify who or what the pronouns and noun phrases, in summary, are referring to. If a person or other entity is mentioned, it should be clear what their role in the story is. So, a reference would be unclear if an entity is referenced but its identity or relation to the story remains unclear.

4. **Focus:** The summary should have a focus; sentences should only contain information that is related to the rest of the summary.

5. **Structure and Coherence:** The summary should be well- structured and well-organized. The summary should not just be a heap of related information but should build from sentence to sentence to a coherent body of information about a topic.

News:
<http://edition.cnn.com/2013/07/04/travel/brazil-10-things/index.html>

Summaries:

1. Brazil might be the biggest country most of the world doesnt know a whole heckuva a lot about. Now that Brazil will be hosting the World Cup next year and the Olympics in 2016 , its time for a crash course in all things Brazilian. In world rankings for the gap between rich and poor, Brazil has the 11th biggest gulf, coming in after a group of impoverished African countries. The seventh largest city in Brazil sits halfway up the Amazon River, where the Rio Negro intersects the great river. In Rio de Janeiro, the city puts on free music fests, with top bands performing on stages across the city, while in Recife, at the citys outdoor concert stage across the street from the beach, free concerts range from rock to forro to an event featuring every drummer with a drum kit in the city, more than 100 of them playing in unison.

2. Brazil might be the biggest country most of the world doesn't know a whole heckuva a lot about. Now that Brazil will be hosting the World Cup next year and the Olympics in 2016 (massive public expenditures that many here deeply oppose), it's time for a crash course in all things Brazilian. Not many Brazilians speak English, particularly outside Rio de Janeiro or Sao Paulo. In world rankings for the gap

between rich and poor, Brazil has the 11th biggest gulf, coming in after a group of impoverished African countries. The seventh largest city in Brazil sits halfway up the Amazon River, where the Rio Negro intersects the great river. Today, the city of Manaus has gleaming office buildings and modern factories and is growing faster than many other parts of the country. The Brazilian knack for improvising on any musical theme makes this country unrivaled in the quantity and quality of different musical rhythms and styles. But Brazil dances to the beat of many different drummers, reflecting its diverse population. Piranha live in the major river basins of South America, and the Amazon tributaries of Brazil are well-stocked with them.

3. Brazil might be the biggest country most of the world doesn't know a whole heckuva a lot about. This is beginning to change as financial fortunes for some have soared in recent years, helping lift some 50% of the country into the middle class. But success hasn't come without problems. Still, if there's one thing Brazil is known for worldwide it's a gregarious and energetic population. Now that Brazil will be hosting the World Cup next year and the Olympics in 2016 (massive public expenditures that many here deeply oppose), it's time for a crash course in all things Brazilian. Brazilians like to express themselves and don't have many restraints on the volume with which they do it. That goes not just for conversation, but car horns, TV pitchmen and sound trucks blasting advertisements through the neighborhood from 16 speakers. They've seen all the American movies and know all the Adele songs, but the premium has traditionally been on getting a good education in Portuguese, still difficult in a land where education is far down the list of priorities. For scale, it's hard to match the immensity of the Amazon rainforest, the largest in the world, or the power of Iguazu Falls, which is the second widest and in volume in the world.

4. Brazil might be the biggest country most of the world doesn't know a whole heckuva a lot about. Now that Brazil will be hosting the World Cup next year and the Olympics in 2016 (massive public expenditures that many here deeply oppose), it's time for a crash course in all things Brazilian. As Portuguese speakers on a Spanish continent far from the English-speaking world, Brazilians have been a linguistic universe unto themselves. Not many Brazilians speak English, particularly outside Rio de Janeiro or Sao Paulo. They've seen all the American movies and know all the Adele songs, but the premium has traditionally been on getting a good education in Portuguese, still difficult in a land where education is far down the list of priorities. Maybe because they're surrounded by so many varieties of exotic fruits the rest of us have never heard of – caju, camu-camu, pitanga – Brazilians are experts in the creation of especially tasty fruit drinks, or sucos. The Brazilian banana is the tastiest

in the world, far superior to the bland Central American version (say Brazilians), and it makes for super savory drinks. Stay in a safe area, don't carry more money on you than you can afford to lose, keep your valuables in the hotel safe, use taxis vetted by your hotel, don't take van taxis and make sure you know what part of town you're in at night. In Rio de Janeiro, the city puts on free music fests, with top bands performing on stages across the city, while in Recife, at the city's outdoor concert stage across the street from the beach, free concerts range from rock to forro to an event featuring every drummer with a drum kit in the city, more than 100 of them playing in unison.

5. ... .10

No total foram aplicados 205 questionários, compostos pelas instruções iniciais, o *link* para o documento origninal e 10 sumários em ordem aleatória e não identificados os sumarizadores, de acordo com o exemplo apresentado acima. Cada questionário foi avaliado por três avaliadores, conforme metodologia descrita na Seção 6.2.1.

Para garantir uma qualidade no processo de avaliação, apenas nove países cuja língua oficial é o inglês foram selecionados para que permitam avaliações. Desses, quatro países tiveram avaliadores. A figura 16 mostra a distribuição do número de avaliadores válidos (alta confiança) por País. O Estados Unidos ficou em primeiro com 63 avaliadores, seguido da Índia com 31 avaliadores, Grã-Bretanha com 13 e Canadá com 10 avaliadores.

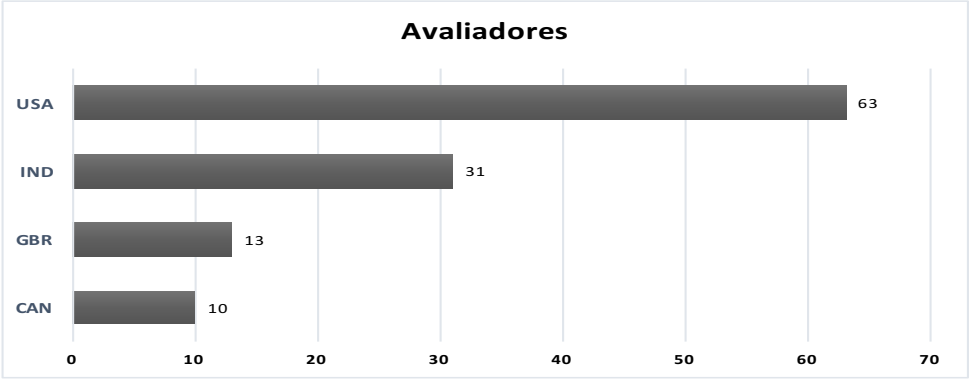


Figura 16 – Distribuição geográfica dos avaliadores.

Durante o processo de avaliação, os avaliadores têm com objetivo indicar o(s) melhor(es) sumário(s) conforme critérios definidos no cabeçalho do questionário de avaliação. Quando a avaliação é concluída, a ferramenta figure-eight agrega os resultados dos três avaliadores com uma pontuação de confiança[6]. O índice de confiança descreve o nível de concordância entre os avaliadores (ponderado pelas pontuações de confiança de cada avaliador) e indica a confiança na validade da resposta agregada para cada questionário. O resultado agregado é escolhido com base na resposta com a maior confiança.

[6] https://success.figure-eight.com/hc/en-us/articles/201855939-How-to-Calculate-a-Confidence-Score

A tabela 46 apresenta os resultados do desempenho geral dos sistemas avaliados de acordo com os resultados agregados da ferramenta *figure eight*. O sistema AutoS obteve os 29 sumários mais bem avaliados, seguido dos sistemas C4J e ALMUS com 25 sumários cada um. Não houve nenhum sumarizador que obtivesse um bom desempenho para todos os documentos sumarizados. Os resultados foram bem distribuídos entre os sumarizadores.

Tabela 46 – Resultados da avaliação manual agregada pela ferramenta Figure Eight. Os sistemas com maior número de sumários que obtiveram o melhor desempenho são destacados em negrito.

Sistema	**#Sumários**
AutoS	**29**
C4J	**25**
ALMUS	**25**
MEAD	22
PLI + GE	19
SUMMA	18
HP-FS	18
TextRank	18
RL + PLI	16
Aylien	15

A tabela 47 apresenta os resultados do desempenho geral dos sistemas avaliados sem agregação da ferramenta *figure eight*. Para selecionar os sumários sem agregação, verificou-se se um sumário foi selecionado por no mínimo dois avaliadores. Assim como na avaliação agregada, o sistema AutoS atingiu o melhor resultado com 52 sumários, seguido dos sistemas C4J com 41 sumários e PLI + GE com 35 sumários.

Tabela 47 – Resultados da avaliação manual sem agregação. Os sistemas com maior número de sumários que obtiveram o melhor desempenho são destacados em negrito.

Sistema	**#Sumários**
AutoS	**52**
C4J	**41**
PLI + GE	**35**
Aylien	33
ALMUS	31
HP-FS	30
SUMMA	29
TextRank	28
MEAD	27
RL + PLI	27

6.2.4 Avaliação comparativa entre as medidas automáticas e avaliação manual

Este último experimento compara o desempenho das medidas automáticas de avaliação de sumários com a avaliação manual. Para esta comparação foi avaliada a intersecção do número de sumários selecionados por cada medida e os selecionados por humanos. Além disso, comparou-se o desempenho das medidas em dois cenários: (i) selecionaram-se os 3 melhores resumos indicados em cada medida versus o resultado da avaliação agregada, para isso, foi analisado se o sumário indicado pela avaliação agregada estava contido nos 3 melhores sumários; e (ii) escolheu-se o melhor resumo indicado por cada medida automática para ver se está contido nos melhores sumários da avaliação não agregada.

A tabela 48 apresentado os resultados obtidos para cada uma das medidas. A medida *Word Concreteness* atingiu o melhor resultado com 38,05% acertando um total de 78 sumários dos 205 avaliados para o primeiro cenário de avaliação. Já para o segundo cenário, a medida *KLSummaryInput* atingiu o melhor resultado com 24,88% de acerto com 51 resumos selecionados corretamente.

As medidas mais utilizadas no estado da arte para avaliação de sumários (ROUGE-1 e ROUGE-2), não atingiram resultados relevantes comparados as avaliações manuais e as demais medidas. No contexto do ROUGE, o R1 atingiu o melhor resultado para os dois cenários de avaliação.

No geral, nenhuma medida conseguiu atingir resultados satisfatórios para seleção automática de sumários para os 205 documentos selecionados, pois não houve correlação satisfatória com a avaliação humana.

Tabela 48 – Resultados da análise comparativa entre as avaliações manuais e as trinta e cinco medidas para avaliação de informatividade e fluência de sumários. Em negrito são destacadas as medidas que atingiram melhor correlação com humanos.

Métodos	**Medidas**	**3 Resumos vs Aval. Agg**		**1 Resumo vs Aval.**	
		#Inter.	**%Acerto**	**#Inter.**	**%Acerto**
Clareza Referencial	RC	73	35,61	46	22,44
Jlnsect	CS	54	26,34	33	16,10
	OS	69	33,66	41	20,00
	SS	71	34,63	42	20,49
	VS	69	33,66	42	20,49
Co-Metrix	Narrativity	61	29,76	31	15,12
	Syntactic Simplicity	76	37,07	30	14,63
	Word Concreteness	**78**	**38,05**	40	19,51
	Referential Cohesion	68	33,17	38	18,54
	Deep Cohesion	61	29,76	38	18,54
	Formality	60	29,27	31	15,12
	Verb Cohesion	64	31,22	32	15,61
	Connectivity	69	33,66	40	19,51
	Temporality	63	30,73	44	21,46
Legibilidade	FleschReadingEase	71	34,63	28	13,66
	FleschKincaidGradeL.	67	32,68	39	19,02
	GunningFogIndex	63	30,73	34	16,59
	ColemanLiauIndex	58	28,29	30	14,63
	SMOGIndex	65	31,71	37	18,05
	AutomatedReadabilityI.	72	35,12	35	17,07
	SpacheScore	61	29,76	43	20,98
	DaleChallScore	62	30,24	40	19,51
	AverageGradeLevel	61	29,76	38	18,54
SiMetrix	KLInputSummary	70	34,15	50	24,39
	KLSummaryInput	65	31,71	**51**	**24,88**
	gunsmoothedJSD	68	33,17	48	23,41
	smoothedJSD	67	32,68	50	24,39
	cosineAllWords	55	26,83	41	20,00
	percentTopicTokens	60	29,27	33	16,10
	fractionTopicWords	66	32,20	46	22,44
	topicWordOverlap	58	28,29	40	19,51
	unigramProb	64	31,22	41	20,00
	multinomialProb	62	30,24	45	21,95
Rouge	R1-F1	75	36,59	42	20,49
	R2-F2	59	28,78	32	15,61

6.3 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este capítulo representa a última contribuição desta tese. Aqui foram apresentadas e analisadas diversas medidas para avaliação automática de informatividade e qualidade textual de sumários extrativos sem o uso de um sumário de referência.

Os resultados experimentais demonstram que: (i) os melhores sistemas avaliados nesta tese não atingiram resultados relevantes para as 35 medidas avaliadas, incluindo as medidas do ROUGE-1 e ROUGE-2; (ii) as medidas para seleção automática de resumo candidato com base na informatividade e fluência não atingiram resultados satisfatórios se comparadas as avaliações manuais; e (iii) a metodologia de avaliação é consistente e, demostra que não

existe uma única medida possível que indique a qualidade textual de um sumário.

No próximo capítulo serão apresentadas as conclusões a respeito do trabalho desenvolvido, suas limitações e as contribuições obtidas, bem como discutidas algumas perspectivas de trabalhos futuros.

7 CONSIDERAÇÕES FINAIS

Neste trabalho de doutorado foi desenvolvida uma abordagem para sumarização automática semi-extrativa e monodocumento de artigos de notícias escritos em inglês. A arquitetura geral da abordagem proposta é composta de cinco etapas principais: (i) A geração de resumos candidatos por sumarizadores extrativos do estado da arte; (ii) a seleção do resumo mais informativo e com maior qualidade textual; (iii) a resolução de anáforas pronominais; (iv) a geração de pronomes; e, por fim, (v) a redução de sentenças. Cada uma dessas etapas tem por objetivo pós-processar um resumo extrativo, gerando assim um novo resumo semi-extrativo mais coeso, legível, gramaticalmente correto e informativo. Além disso, as abordagens propostas não são supervisionadas, ou seja, não necessitam de conhecimento prévio acerca dos problemas apresentados nesta tese.

As abordagens propostas foram avaliadas adotando os principais corpora das áreas de sumarização monodocumento de artigos de notícias escritos em inglês: CNN, DUC 2001 e DUC 2002. O Corpus CNN foi construído em paralelo ao desenvolvimento deste trabalho de doutorado. De acordo com o conhecimento do autor desta tese e seu supervisor, esse corpus é o maior corpus de notícias para sumarização monodocumento e extrativa do estado da arte.

Para avaliar os resumos gerados, adotaram-se as principais medidas automáticas de avaliação do estado da arte (as medidas do ROUGE-1, ROUGE-2, similaride do cosseno e interseção de sentenças) e avaliações humanas realizadas nas plataformas do *figure-eight* e *Amazon Mechanical Turk*. Além disso, para a avaliação automática da qualidade dos sumários foram exploradas 33 medidas que independem de um resumo de referência. No geral, diversos aspectos de cada uma das propostas desenvolvidas foram avaliados e comparados com o desempenho de referência dos sistemas de sumarização do estado da arte.

Os resultados experimentais obtidos demonstraram que as propostas aqui formuladas apresentam um desempenho competitivo, melhorando o nível de coesão, legibilidade e informatividade dos sistemas do estado da arte selecionados para tarefa de sumarização extrativa monodocumento. Outra descoberta importante é que as 33 medidas de avaliação de sumários, considerando a qualidade textual e a informatividade, não apresentaram correlação estatísticamente relevante com as avaliações humanas. Isso se aplica também, as medidas de avaliação do ROUGE-1 e ROUGE-2.

Portanto, esta pesquisa de doutorado alcançou uma série de resultados na área. Sendo as maiores contribuições na área de geração de sumários coesos. Acredita-se que por meio da sumarização semi-extrativa é possível atingir bons resultados para melhoria da qualidade dos resumos.

7.1 CONTRIBUIÇÕES DO TRABALHO

Dentre as contribuições alcançadas neste trabalho de doutorado, podem-se enumerar:

1. Desenvolvimento de uma metodologia para criação do *Corpus CNN*, mapeando os highlights nas sentenças da notícia. O *corpus* é de fácil entendimento e leitura, escolhendo-se notícias do portal CNN, tal *corpus* foi coletado, processado, revisado e obtidos sumários abstrativos sugeridos pelos autores, chamados de *highlights*, chegando a 3.000 documentos; Criaram-se sumários extrativos de referência (*gold standard*) para todos os documentos, escolhidos por um grupo de especialistas, os quais foram classificados sumários de confiança para fins de avaliação; a escolha desse *corpus* é evidenciada pela facilidade de entendimento do conteúdo frente à concorrência.

2. Avaliação quantitativa e qualitativa de sistemas de sumarização automática de texto. Para isso foi necessário automatizar 22 sistemas comerciais e gratuitos, a fim de avaliar o desempenho de cada um deles em um grande *corpus* de sumarização. Na etapa de experimentos, foram gerados 66 mil sumários para avaliação quantitativa dos sistemas selecionados.

3. Desenvolvimento de um enfoque independente de sistema de sumarização para resolução de anáforas pronominais, contendo:

   (i) um componente para identificação de expressões anafóricas em sumários extrativos, que possibilitou explorar um dos principais problemas de coesão em sumários extrativos. Além disso, esse componente poderá ser usado para avaliar o nível de qualidade de um sumário a partir do número de pronomes não conectados, ou seja, sem uma entidade (por exemplo, um nome) ao qual ele se refere;

   (ii) um componente para resolução de anáforas pronominais no texto-fonte (pré-processamento), gerando a partir dele um texto intermediário. Esse texto, com as correferências resolvidas, poderá ser usado por qualquer sumarizador extrativo, assim as sentenças selecionadas serão autocontidas (todos os pronomes conectados).

   (iii) um componente para resolução de anáforas pronominais no sumário extrativo. Esse componente foi inserido na etapa de pós-processamento dos sumários, gerando assim uma nova versão semi-extrativa com as correferências quebradas corrigidas. Para validação da abordagem proposta, foram gerados automaticamente 189 mil sumários durante a etapa de experimentos.

4. Concepção de uma abordagem para geração e reinserção de pronomes em sumários extrativos, gerando assim novos sumários semi-extrativos mais coesos. A solução

proposta aplica o método desenvolvido nesta tese para geração de cadeias de correferências pronominais, em seguida mapeia as entidades repetidas nos resumos e sustitui elas por pronomes. Durante a avaliação da abordagem proposta constatou-se que todos os sistemas do estado da arte selecionados geram sumários com entidades repetidas e para validar a proposta foram usados métodos de avaliação automática e manual.

5. Desenvolvimento de uma abordagem baseada em regras para simplificação de sentenças em tarefa de sumarização, sem perder a informatividade e a legibilidade do resumo gerado. Tais regras foram avaliadas automaticamente e por humanos.

6. Realização de uma avaliação exaustiva de diversas medidas para estimar a informatividade, legibilidade e coesão de um texto, a fim de selecionar o melhor resumo candidato. A arquitetura da abordagem proposta é dividida em duas etapas centrais: (i) geração dos resumos candidatos; e (ii) seleção do resumo mais informativo e qualitativo. A geração dos candidatos a resumo é executada pelos melhores sistemas de sumarização do estado da arte selecionados nesta tese.

7.2 PRODUÇÃO BIBLIOGRÁFICA

Esta seção apresenta as listas de artigos produzidos durante o desenvolvimento deste trabalho de doutorado. A primeira mostra os artigos publicados e a segunda os submetidos.

Artigos publicados:

1. **Jamilson Batista**, Rafael Dueire Lins, Rinaldo Lima, Hilário Oliveira, Marcelo Riss, Steven J. Simske, 2018. Automatic Cohesive Summarization with Pronominal Anaphora Resolution. Journal "Computer Speech & Language", p. –, 2018. ISSN 0885-2308. DOI: <https://doi.org/10.1016/j.csl.2018.05.004>

2. **Jamilson Batista Antunes**, Rafael Dueire Lins, Rinaldo Lima, Steven J. Simske and Marcelo Riss, 2016. Towards Cohesive Extractive Summarization through Anaphoric Expression Resolution. In Proceedings of the 2016 ACM Symposium on Document Engineering (DocEng '16). ACM, Vienna, Austria. DOI: <http://dx.doi.org/10.1145/2960811.2967159>

3. **Jamilson Batista**, Rodolfo Ferreira, Hilário Tomaz, Rafael Ferreira, Rafael Dueire Lins, Steven Simske, Gabriel Silva, and Marcelo Riss, 2015. A Quantitative and Qualitative Assessment of Automatic Text Summarization Systems. In Proceedings of the 2015 ACM Symposium on Document Engineering (DocEng '15). ACM, New York, NY, USA, 65-68. DOI<http://dx.doi.org/10.1145/2682571.2797081>

4. Rinaldo J. Lima, **Jamilson Batista**, Rafael Ferreira, Fred Freitas, Rafael Dueire Lins, Steven Simske, and Marcelo Riss, 2014. Transforming graph-based sentence

representations to alleviate overfitting in relation extraction. In Proceedings of the 2014 ACM symposium on Document engineering (DocEng '14). ACM, New York, NY, USA, 53-62. DOI<http://dx.doi.org/10.1145/2644866.2644875>

Artigo submetido:

1. Rafael Dueire Lins, Hilário Oliveira, Bruno T. Ávila, Luciano Cabral, **Jamilson Antunes**, Diego A. Salcedo, Rafael Ferreira, Gabriel F. Silva, Rinaldo Lima and Steven J. Simske, 2018. The CNN-Corpus: A Large Textual Corpus for Single-Document Extractive Summarization. Journal "Language Resources and Evaluation".

7.3 LIMITAÇÕES

As principais limitações das abordagens propostas neste trabalho são: **(i) a sua natureza monodocumento**, ou seja, os resumos são criados a partir de um único documento de entrada; **(ii) o viés com o tipo dos documentos de entrada a serem sintetizados**. Este trabalho teve como escopo, somente a sumarização de artigos de notícias. Apesar das abordagens propostas neste trabalho (resolução de anáfora, geração de pronomes, redução de sentenças e seleção de sumários) serem independentes do tipo de documento de entrada, não existe nenhuma garantia de que elas também apresentem um bom desempenho para outros tipos de documentos textuais, tais como artigos científicos, e-mails, *blogs*, entre outros; e **(iii) a seleção de sumários independente de um sumário de referência.** As medidas usadas neste trabalho foram avaliadas individualmente e apesar delas não apresentarem resultados satisfatórios, não foi avaliada a combinação entre elas, por exemplo usando um método de regressão.

As abordagens propostas foram avaliadas indivualmente, com isso representam uma importante limitação, pois as abordagens não foram avaliadas de forma integrada, isto é, em um único fluxo.

7.4 TRABALHOS FUTUROS

Existem vários desafios abertos que podem ser abordados no futuro, a fim de prosseguir com as linhas de pesquisa exploradas nesta tese. 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:

- **Combinação de sumários.** Geração de um sumário híbrido por meio de duas estratégias propostas e testadas em CABRAL (2015): (I) votação (sentenças comuns selecionadas por no mínimo 50% dos sumarizadores); e (II) uma variante do algoritmo *K-means*. Essa variante foi usada para descartar as sentenças, caso exceda a taxa de compressão. Ela é baseada na medida *Average_R* fornecida pelo ROUGE para

cada um dos sumarizadores avaliados, a fim de selecionar as sentenças de maior pontuação. O autor demostrou que um sumário formado a partir da fusão de vários sumários gerados por diferentes ferramentas apresentou sucesso de acordo com o ROUGE, e tal estratégia parece ser muito promissora em termos de maximizar o nível de melhoria quantitativa da sumarização.

A desvantagem da abordagem proposta por CABRAL (2015) é que o processo é manual (mínimo de três leitores qualificados), dependente de um sumário extrativo de referência (*gold stantard*) para avaliação do ROUGE e não possui um mecanismo para evitar redundância de sentenças muito similares. Em contribuição, o presente trabalho se propõe a usar a média global das medidas de avaliação descritas no Capítulo 6, em vez do ROUGE, e a usar a medida de similaridade desenvolvida por Ferreira et al. (2016) para remover sentenças redundantes, ou seja, com alta similaridade entre si.

- **Fusão de sentenças.** Pesquisas envolvendo o desenvolvimento de abordagens de sumarização abstrativas têm crescido nos últimos anos. Esses sistemas adotam, principalmente, técnicas de fusão de sentenças. Banerjee, Mitra e Sugiyama (2015) propuseram 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 fusão de sentenças têm apresentado bom desempenho, gerando resumos menores e mais informativos do que sistemas de sumarização extrativos.

- **Adaptação das abordagens propostas para artigos de notícias escritas em português do Brasil.** Pesquisas envolvendo documentos escritos em português do Brasil ainda são escassas em relação a outros idiomas, como o inglês. Diante disso, vislumbra-se adaptar as abordagens propostas para realizar o processo de sumarização semi-extrativa e monodocumento, em artigos de notícias escritos em português. Para avaliar as adaptações realizadas, experimentos podem ser realizados adotando o corpus CST-News (ALEIXO; PARDO, 2008; CARDOSO et al., 2011), que é muito utilisado em pesquisas envolvendo artigos de notícias escritas em português.

- **Sumarização multidocumento.** Um dos principais focos desta pesquisa é a sumarização monodocumento, pois ainda apresenta resultados satisfatórios no estado da arte, porém pretende-se no futuro adaptar as estratégias propostas nesta tese para sumarização multidocumento, pois nessa área de pesquisa têm-se problemas mais recorrêntes de correferência em aberto, redudância de sentenças, falta de legibilidade, entre outros.

- **Combinação de medidas de avaliação.** Neste trabalho de doutorado foram usadas 35 medidas de avaliação de sumários e nenhuma delas atingiu resultados

satisfatórios comparadas a avaliação humana, porém Ellouze, Jaoua e Belguith (2017) mostraram em seus experimentos que a combinação de algumas medidas de avaliação tem desmonstrado bons resultados. Portanto, pretende-se no futuro realizar a combinação das medidas usadas nesse trabalho a fim de selecionar o melhor sumário informativo e qualitativo.

- **Teste integrado.** Nesta tese foram avaliadas individualmente cada abordagem proposta. Pretende-se como trabalho futuro testar de forma integradar todas as abordagens aqui propostas.

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.

ALEIXO, P.; PARDO, T. A. S. *CSTNews: Um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (cross-document structure theory*. [S.l.], 2008.

ALGULIEV, R. M.; ALIGULIYEV, R. M.; HAJIRAHIMOVA, M. S.; MEHDIYEV, C. A. Mcmr: Maximum coverage and minimum redundant text summarization model. *Expert Systems with Applications*, v. 38, n. 12, p. 14514 – 14522, 2011. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417411008177>.

AUTOSUMMARIZER. *Automatic Text Summarizer*. 2014. Http://autosummarizer.com/. Last acess: Mar. 2015.

AYLIEN. *Aylien Text Analysis API*. 2011. Http://aylien.com/text-api-doc. Last acess: Mar. 2015.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern Information Retrieval*. 1st. ed. [S.l.]: Addison Wesley, 1999. Paperback. ISBN 020139829X.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999. ISBN 020139829X.

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. Disponível em: <http://doi.acm.org/10.1145/2740908.2742751>.

BARRERA, A.; VERMA, R. Combining syntax and semantics for automatic extractive single-document summarization. In: *Proceedings of the 13th international conference on Computational Linguistics and Intelligent Text Processing - Volume Part II*. Berlin, Heidelberg: Springer-Verlag, 2012. (CICLing'12), p. 366–377. ISBN 978-3-642-28600-1. Disponível em: <http://dx.doi.org/10.1007/978-3-642-28601-8_31>.

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 2015 ACM Symposium on Document Engineering*. New York, NY, USA: ACM, 2015. (DocEng '15), p. 65–68. ISBN 978-1-4503-3307-8.

BATISTA, J.; LINS, R. D.; LIMA, R.; OLIVEIRA, H.; RISS, M.; SIMSKE, S. J. Automatic cohesive summarization with pronominal anaphora resolution. *Computer Speech & Language*, p. –, 2018. ISSN 0885-2308. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0885230816302340>.

BIBER, D. *Variation across Speech and Writing*. [S.l.]: Cambridge University Press, 1988.

BIBER, D.; CONRAD, S.; REPPEN RANDI, R. *Corpus Linguistics: Investigating Language Structure and Use*. New York, NY, USA: Cambridge University Press, 1998. ISBN 9780521499576.

BOSCH, A. van den; BOGERS, T.; KUNDER, M. de. Estimating search engine index size variability: a 9-year longitudinal study. *Scientometrics*, v. 107, n. 2, p. 839–856, 2016. ISSN 1588-2861. Disponível em: <http://dx.doi.org/10.1007/s11192-016-1863-z>.

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. Disponível em: <http://aclweb.org/anthology/D15-1220>.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, v. 30, n. 1–7, p. 107–117, 1998.

BURNARD, L. *Users Reference Guide British National Corpus Version 1.0*. [S.l.], 1995.

CABRAL, L.; LIMA, R.; LINS, R.; NETO, M.; FERREIRA, R.; SIMSKE, S.; RISS, M. Automatic summarization of news articles in mobile devices. In: *2015 Fourteenth Mexican International Conference on Artificial Intelligence (MICAI)*. [S.l.: s.n.], 2015. p. 8–13.

CABRAL, L. D. S. *Uma Plataforma para Sumarização Automática de Textos Independente de Idioma*. Tese (Doutorado) — Universidade Federal de Pernambuco, 2015. Disponível em: <http://repositorio.ufpe.br/handle/123456789/14968>.

CABRAL, L. S.; LINS, R. D.; MELLO, R. F.; FREITAS, F.; ÁVILA, B.; SIMSKE, S.; RISS, M. A platform for language independent summarization. In: *Proceedings of the 2014 ACM Symposium on Document Engineering*. New York, NY, USA: ACM, 2014. (DocEng '14), p. 203–206. ISBN 978-1-4503-2949-1.

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*. AAAI Press, 2015. (AAAI'15), p. 2153–2159. ISBN 0-262-51129-0. Disponível em: <http://dl.acm.org/citation.cfm?id=2886521.2886620>.

CARDOSO, P. C. F.; MAZIERO, E. G.; JORGE, M. L. del R. C.; ELOIZE, M.; SENO, R.; FELIPPO, A. D.; RINO, L. H. M.; NUNES, M. das G. V.; PARDO, T. A. S. Cstnews-a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In: . [S.l.: s.n.], 2011.

CARRELL, P. L. Cohesion is not coherence*. *TESOL Quarterly*, Blackwell Publishing Ltd, v. 16, n. 4, p. 479–488, 1982. ISSN 1545-7249. Disponível em: <http://dx.doi.org/10.2307/3586466>.

CHRISTENSEN, J. *Towards Large Scale Summarization*. Tese (Doutorado) — University of Washington, Seattle, WA, USA, 2014.

CHRISTENSEN, J.; SODERL, S.; ETZIONI, O. Towards coherent multi-document summarization. In: *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. [S.l.: s.n.], 2013. (NAACL 2013).

COLEMAN, M.; LIAU, T. L. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, v. 60, p. 283–284, 1975.

COLLOVINI, S.; CARBONEL, T.; FUCHS, J. T.; COELHO, J. C.; RINO, L. H. M.; VIEIRA, R. Summit: Um corpus anotado com informações discursivas visando à sumarização automática. In: *Proceedings of the SBC, Workshop em Tecnologia da Informação e da Linguagem Humana (TIL'2007), Rio de Janeiro, RJ*. [S.l.: s.n.], 2007.

COMPACTOR, T. *Text Compactor*. 2015. Http://www.textcompactor.com. Last acess: Mar. 2015.

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.

CRYSTAL, D. *Style: the varieties of English*. 2nd. ed. [S.l.]: Penguin Books, 1987. 241–263 p. (A history of literature in the English language).

CUNNINGHAM, H.; MAYNARD, D.; BONTCHEVA, K.; TABLAN, V. Gate: An architecture for development of robust hlt applications. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. (ACL '02), p. 168–175. Disponível em: <https://doi.org/10.3115/1073083.1073112>.

DALE, E.; CHALL, J. *A Formula for Predicting Readability*. Bureau of Educational Research, Ohio State University, 1948. Disponível em: <https://books.google.com.br/books?id=kxgAGQAACAAJ>.

DALIANIS, H.; AL. et. From swesum to scandsum: Automatic text summarization for the scandinavian languages. In: HOLMBOE, H. (Ed.). *ScandSum*. Museum Tusculanums Forlag, 2003. p. 153–163. Disponível em: <http://nlp.lacasahassel.net/publications/scandsum02.pdf>.

DANG, H. T. Overview of DUC 2005. In: *Proceedings of the Document Understanding Conference*. [S.l.: s.n.], 2005.

DAS, D.; MARTINS, A. F. T. *A Survey on Automatic Text Summarization*. [S.l.], 2007.

DONAWAY, R. L.; DRUMMEY, K. W.; MATHER, L. A. A comparison of rankings produced by summarization evaluation measures. In: *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic Summarization - Volume 4*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000. (NAACL-ANLP-AutoSum '00), p. 69–78.

DORR, B. J.; ZAJIC, D. M.; SCHWARTZ, R. M. Hedge trimmer: A parse-and-trim approach to headline generation. In: . [S.l.: s.n.], 2003.

DOWELL, N.; GRAESSER, A.; CAI, Z. Language and discourse analysis with coh-metrix: Applications from educational material to learning environments at scale. *Journal of Learning Analytics*, 2015.

DOWELL, N.; GRAESSER, A.; CAI, Z. Language and discourse analysis with coh-metrix: Applications from educational material to learning environments at scale. v. 3, 04 2015.

DUNLAVY, D. M.; O'LEARY, D. P.; CONROY, J. M.; SCHLESINGER, J. D. Qcs: A system for querying, clustering and summarizing documents. *Information Processing & Management*, v. 43, n. 6, p. 1588 – 1605, 2007. ISSN 0306-4573. Text Summarization. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0306457307000246>.

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 (Volume 1: Long Papers)*. Association for Computational Linguistics, 2016. p. 1998–2008. Disponível em: <http://www.aclweb.org/anthology/P16-1188>.

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. Disponível em: <http://doi.acm.org/10.1145/321510.321519>.

ELLOUZE, S.; JAOUA, M.; BELGUITH, L. Merging multiple features to evaluate the content of text summary. v. 58, p. 69–76, 03 2017.

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. Disponível em: <http://dl.acm.org/citation.cfm?id=1622487.1622501>.

EVANS, R. Applying machine learning toward an automatic classification of *It. LLC*, v. 16, n. 1, p. 45–57, 2001.

FATTAH, M. A.; REN, F. Ga, mr, ffnn, pnn and gmm based models for automatic text summarization. *Comput. Speech Lang.*, Academic Press Ltd., London, UK, UK, v. 23, n. 1, p. 126–144, jan. 2009. ISSN 0885-2308. Disponível em: <http://dx.doi.org/10.1016/j.csl.2008.04.002>.

FERREIRA, R.; CABRAL, L. S.; FREITAS, F. L. G.; LINS, R. D.; SILVA, G. F. Pereira e; SIMSKE, S. J.; FAVARO, L. A multi-document summarization system based on statistics and linguistic treatment. *Expert Syst. Appl.*, v. 41, n. 13, p. 5780–5787, 2014.

FERREIRA, R.; CABRAL, L. S.; LINS, R. D.; SILVA, G. Pereira e; FREITAS, F.; CAVALCANTI, G. D. C.; 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.; CABRAL, L. S.; LINS, R. D.; LIMA, R.; SILVA, G. F. 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.

FERREIRA, R.; LINS, R. D.; SIMSKE, S. J.; FREITAS, F.; RISS, M. Assessing sentence similarity through lexical, syntactic and semantic analysis. *Computer Speech & Language*, v. 39, p. 1 – 28, 2016. ISSN 0885-2308. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0885230816000048>.

FILIPPOVA, K. Multi-sentence compression: Finding shortest paths in word graphs. In: *COLING'10*. [S.l.: s.n.], 2010. p. 322–330.

FILIPPOVA, K.; ALTUN, Y. Overcoming the lack of parallel data in sentence compression. p. 1481–1491, 01 2013.

FINDWISE. *Findwise Multi-document Summarizers*. 2015. Http://labdemos.findwise.com/demomds. Last acess: Mar. 2015.

FORBES, C.; EVANS, M.; HASTINGS, N.; PEACOCK, B. *Statistical Distributions*. Wiley, 2011. ISBN 9781118097823. Disponível em: <https://books.google.com.br/books?id=YhF1osrQ4psC>.

FREE Summarizer. 2011. Http://freesummarizer.com/. Last acess: Mar. 2015.

GAMBHIR, M.; GUPTA, V. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, Springer Netherlands, 2016. ISSN 0269-2821. Disponível em: <http://link.springer.com/10.1007/s10462-016-9475-9>.

GARCÍA-HERNÁNDEZ, R. A.; LEDENEVA, Y. Single extractive text summarization based on a genetic algorithm. In: CARRASCO-OCHOA, J. A.; MARTÍNEZ-TRINIDAD, J. F.; RODRÍGUEZ, J. S.; BAJA, G. S. di (Ed.). *Pattern Recognition*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 374–383.

GIANNAKOPOULOS, G.; KARKALETSIS, V. Autosummeng and memog in evaluating guided summaries. In: *TAC*. [S.l.: s.n.], 2011.

GIANNAKOPOULOS, G.; KARKALETSIS, V.; VOUROS, G.; STAMATOPOULOS, P. Summarization system evaluation revisited: N-gram graphs. *ACM Trans. Speech Lang. Process.*, ACM, New York, NY, USA, v. 5, n. 3, p. 5:1–5:39, out. 2008. ISSN 1550-4875. Disponível em: <http://doi.acm.org/10.1145/1410358.1410359>.

GIANNAKOPOULOS, G.; VOUROS, G. A.; KARKALETSIS, V. MUDOS-NG: multi-document summaries using n-gram graphs (tech report). *CoRR*, abs/1012.2042, 2010. Disponível em: <http://arxiv.org/abs/1012.2042>.

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.

GIVÓN, T. *Functionalism and Grammar*. J. Benjamins, 1995. ISBN 9789027221476. Disponível em: <https://books.google.com.br/books?id=1KWuyR4R-kUC>.

GONÇALVES, P. N.; RINO, L. H. M.; VIEIRA, R. Summarizing and referring: towards cohesive extracts. In: *Proceedings of the 2008 ACM Symposium on Document Engineering, Sao Paulo, Brazil, September 16-19, 2008*. [S.l.: s.n.], 2008. p. 253–256.

GONG, Y.; LIU, X. Generic text summarization using relevance measure and latent semantic analysis. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2001. (SIGIR '01), p. 19–25. ISBN 1-58113-331-6. Disponível em: <http://doi.acm.org/10.1145/383952.383955>.

GRAESSER, A. C.; MCNAMARA, D.; LOUWERSE, M. What readers need to learn in order to process coherence relations in narrative and expository text. In: *Rethinking Reading Comprehension*. [S.l.: s.n.], 2003. p. 82–.

GRAESSER, A. C.; MCNAMARA, D. S. Computational analyses of multilevel discourse comprehension. *Topics in cognitive science*, v. 3 2, p. 371–98, 2011.

GRAESSER, A. C.; MCNAMARA, D. S.; CAI, Z.; CONLEY, M.; LI, H.; PENNEBAKER, J. Coh-metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal*, v. 115, n. 2, p. 210–229, 2014.

GRAESSER, A. C.; MCNAMARA, D. S.; KULIKOWICH, J. M. Coh-metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, v. 40, n. 5, p. 223–234, 2011.

GRICE, H. P. Logic and conversation. In: *Syntax and Semantics: Vol. 3: Speech Acts*. San Diego, CA: Academic Press, 1975. p. 41–58.

GUINAUDEAU, C.; STRUBE, M. Graph-based local coherence modeling. In: *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. [S.l.: s.n.], 2013. v. 1, p. 93–103.

GUNNING, R. *The Technique of Clear Writing*. McGraw-Hill, 1952. Disponível em: <https://books.google.com.br/books?id=ofI0AAAAMAAJ>.

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. Disponível em: <http://dl.acm.org/citation.cfm?id=1620754.1620807>.

HAHN, U.; MANI, I. The challenges of automatic summarization. *Computer*, IEEE Computer Society Press, Los Alamitos, CA, USA, v. 33, n. 11, p. 29–36, nov. 2000. ISSN 0018-9162. Disponível em: <http://dx.doi.org/10.1109/2.881692>.

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*. Washington, DC, USA: IEEE Computer Society, 2010. (IALP '10), p. 257–260. ISBN 978-0-7695-4288-1. Disponível em: <http://dx.doi.org/10.1109/IALP.2010.45>.

HASLER, L.; ORăSAN, C.; MITKOV, R. Building better corpora for summarisation. In: *Proceedings of Corpus Linguistics 2003*. Lancaster, UK: [s.n.], 2003. p. 309 – 319.

HASSEL, M.; DALIANIS, H. *SweSum - Automatic Text Summarizer*. 2003. Http://swesum.nada.kth.se/index-eng.html. Last acess: Mar. 2015.

HEYLIGHEN, F. Advantages and limitations of formal expression. *Foundations of Science*, v. 4, n. 1, p. 25–56, Mar 1999. ISSN 1572-8471. Disponível em: <https://doi.org/10.1023/A:1009686703349>.

HEYLIGHEN, F.; DEWAELE, J.-M. Variation in the contextuality of language: An empirical measure. *Foundations of Science*, v. 7, n. 3, p. 293–340, Sep 2002. Disponível em: <https://doi.org/10.1023/A:1019661126744>.

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*. Association for Computational Linguistics, 2013. p. 1515–1520. Disponível em: <http://www.aclweb.org/anthology/D13-1158>.

HONG, K.; CONROY, J.; 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)*. [S.l.]: European Language Resources Association (ELRA), 2014.

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*. Lisbon, Portugal: Association for Computational Linguistics, 2015. p. 107–117. Disponível em: <http://aclweb.org/anthology/D15-1011>.

HOVY, E. Text summarization. In: MITKOV, R. (Ed.). *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press, 2004, (Oxford Handbooks in Linguistics). cap. 32, p. 583–598. Disponível em: <http://www.isi.edu/natural-language/people/hovy/papers/05Handbook-Summ-hovy.pdf>.

HOVY, E.; LIN, C.-Y. *Automated Text Summarization in SUMMARIST*. 1999.

HUTCHINS, J. Summarization: Some problems and methods. In: *Meaning: The Frontier of Informatics*. [S.l.]: Aslib, 1987. p. 151–173.

INTELLEXER Summarizer. 2011. Http://summarizer.intellexer.com/. Last acess: Mar. 2015.

JONES, K. *Discourse modelling for automatic summarising*. University of Cambridge, Computer Laboratory, 1993. (Technical report (University of Cambridge. Computer Laboratory)). Disponível em: <https://books.google.com.br/books?id=hHsEAQAAIAAJ>.

JONES, K. S. Automatic summarising: Factors and directions. In: *Advances in Automatic Text Summarization*. [S.l.]: MIT Press, 1998. p. 1–12.

JONES, K. S. Automatic summarising: The state of the art. *Information Processing Management*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 43, n. 6, p. 1449–1481, nov. 2007. ISSN 0306-4573. Disponível em: <http://dx.doi.org/10.1016/j.ipm.2007.03.009>.

KASHANI, M. M.; POPOWICH, F. Pronoun generation for text summarization and question answering. *Proceedings of 5th Slovenian and 1st international Language Technologies Conference*, v. 2006, 2006. Disponível em: <http://nl.ijs.si/is-ltc06/proc/16{_}Kashani.>

KASPERSSON, T.; SMITH, C.; DANIELSSON, H.; JöNSSON, A. This also affects the context - errors in extraction based summaries. In: CHAIR), N. C. C.; CHOUKRI, K.; DECLERCK, T.; DOğAN, M. U.; MAEGAARD, B.; MARIANI, J.; MORENO, A.; ODIJK, J.; PIPERIDIS, S. (Ed.). *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), 2012. ISBN 978-2-9517408-7-7.

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: *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*. [S.l.: s.n.], 2014. v. 2, p. 315–320.

KINCAID, J. P.; FISHBURNE, R. P.; ROGERS, R. L.; CHISSOM, B. S. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. [S.l.], 1975. Disponível em: <http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED108134>.

KULKARNI, U. V.; PRASAD, R. S. Implementation and evaluation of evolutionary connectionist approaches to automated text summarization. In: *Journal of Computer Science*. [S.l.]: Science Publications, 2010.

KULLBACK, S.; LEIBLER, R. A. On information and sufficiency. *Ann. Math. Statist.*, The Institute of Mathematical Statistics, v. 22, n. 1, p. 79–86, 03 1951. Disponível em: <https://doi.org/10.1214/aoms/1177729694>.

LABOV, W. *Sociolinguistic Patterns*. University of Pennsylvania Press, Incorporated, 1972. (Conduct and Communication). ISBN 9780812210521. Disponível em: <https://books.google.com.br/books?id=hD0PNMu8CfQC>.

LANDAUER, T. K.; DUMAIS, S. T. A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, v. 104, p. 211–240, 1997.

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. Disponível em: <http://dx.doi.org/10.1162/COLI_a_00152>.

LEE, H.; PEIRSMAN, Y.; CHANG, A.; CHAMBERS, N.; SURDEANU, M.; JURAFSKY, D. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. (CONLL Shared Task '11), p. 28–34. ISBN 9781937284084.

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*. The Association for Computational Linguistics, 2015. p. 778–787. ISBN 978-1-941643-49-5. Disponível em: <http://dblp.uni-trier.de/db/conf/naacl/naacl2015.html#0013LZ15>.

LIMA, R. J.; BATISTA, J.; FERREIRA, R.; FREITAS, F.; LINS, R. D.; SIMSKE, S.; RISS, M. Transforming graph-based sentence representations to alleviate overfitting in relation extraction. In: *Proceedings of the 2014 ACM Symposium on Document Engineering*. New York, NY, USA: ACM, 2014. (DocEng '14), p. 53–62. ISBN 978-1-4503-2949-1. Disponível em: <http://doi.acm.org/10.1145/2644866.2644875>.

LIN, C.-Y. Rouge: A package for automatic evaluation of summaries. In: MOENS, M.-F.; SZPAKOWICZ, S. (Ed.). *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 74–81.

LIN, C.-Y.; CAO, G.; GAO, J.; NIE, J.-Y. An information-theoretic approach to automatic evaluation of summaries. In: *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006. (HLT-NAACL '06), p. 463–470. Disponível em: <http://dx.doi.org/10.3115/1220835.1220894>.

LIN, C.-Y.; HOVY, E. The automated acquisition of topic signatures for text summarization. In: *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000. (COLING '00), p. 495–501. ISBN 1-55860-717-X. Disponível em: <http://dx.doi.org/10.3115/990820.990892>.

LINS, R. D.; OLIVEIRA, H.; AVILA, B. T.; CABRAL, L.; ANTUNES, J.; SALCEDO, D. A.; FERREIRA, R.; SILVA, G. F.; LIMA, R.; SIMSKE, S. J. The cnn-corpus: A large textual corpus for single-document extractive summarization. *Review of Language Resources and Evaluation*, Springer, 2018.

LINS, R. D.; SIMSKE, S. J.; CABRAL, L. de S.; SILVA, G. de F.; LIMA, R.; MELLO, R. F.; FAVARO, L. A multi-tool scheme for summarizing textual documents. In: *Proc. of 11st IADIS International Conference WWW/INTERNET 2012*. [S.l.: s.n.], 2012. p. 1–8.

LIU, X.; WEBSTER, J. J.; KIT, C. An extractive text summarizer based on significant words. In: *Proceedings of the 22nd International Conference on Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy*. Berlin, Heidelberg: Springer-Verlag, 2009. (ICCPOL '09), p. 168–178. ISBN 978-3-642-00830-6. Disponível em: <http://dx.doi.org/10.1007/978-3-642-00831-3_16>.

LLORET, E.; PALOMAR, M. A gradual combination of features for building automatic summarization systems. In: *Proceedings of the 12th International Conference on Text, Speech and Dialogue*. Berlin, Heidelberg: Springer-Verlag, 2009. (TSD '09), p. 16–23. ISBN 978-3-642-04207-2. Disponível em: <http://dx.doi.org/10.1007/978-3-642-04208-9_6>.

LLORET, E.; PALOMAR, M. Text summarisation in progress: a literature review. *Artif. Intell. Rev.*, Kluwer Academic Publishers, Norwell, MA, USA, v. 37, n. 1, p. 1–41, jan. 2012. ISSN 0269-2821. Disponível em: <http://dx.doi.org/10.1007/s10462-011-9216-z>.

LLORET, E.; PALOMAR, M. Compendium: a text summarisation tool for generating summaries of multiple purposes, domains, and genres. *Natural Language Engineering*, v. 19, p. 147–186, 4 2013. ISSN 1469-8110. Disponível em: <http://journals.cambridge.org/article_S1351324912000198>.

LOUIS, A. Automatic metrics for genre-specific text quality. In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. (NAACL HLT '12), p. 54–59. ISBN 978-1-937284-20-6. Disponível em: <http://dl.acm.org/citation.cfm?id=2385736.2385748>.

LOUIS, A.; NENKOVA, A. Automatic summary evaluation without human models. In: *TAC*. NIST, 2008. Disponível em: <http://dblp.uni-trier.de/db/conf/tac/tac2008.html#LouisN08>.

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. Disponível em: <http://dl.acm.org/citation.cfm?id=1699510.1699550>.

LOUIS, A.; NENKOVA, A. Predicting summary quality using limited human input. In: *Proceedings of the Second Text Analysis Conference, TAC 2009, Gaithersburg, Maryland, USA, November 16-17, 2009*. [s.n.], 2009. Disponível em: <http://www.nist.gov/tac/publications/2009/participant.papers/UPenn.proceedings.pdf>.

LOUIS, A.; NENKOVA, A. Automatically assessing machine summary content without a gold standard. *Comput. Linguist.*, MIT Press, Cambridge, MA, USA, v. 39, n. 2, p. 267–300, jun. 2013. ISSN 0891-2017. Disponível em: <http://dx.doi.org/10.1162/COLI_a_00123>.

LOUIS, A.; NENKOVA, A. What makes writing great? first experiments on article quality prediction in the science journalism domain. *Transactions of the Association for Computational Linguistics*, v. 1, p. 341–352, 2013. ISSN 2307-387X. Disponível em: <https://www.transacl.org/ojs/index.php/tacl/article/view/76>.

LOUIS, A. P. *Predicting text quality: Metrics for content, organization and reader interest*. Tese (Doutorado) — University of Pennsylvania, 2013. Disponível em: <http://repository.upenn.edu/dissertations/AAI3564417/>.

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. Disponível em: <http://dx.doi.org/10.1147/rd.22.0159>.

MACKIE, S.; MCCREADIE, R.; MACDONALD, C.; OUNIS, I. On choosing an effective automatic evaluation metric for microblog summarisation. In: *Proceedings of the 5th Information Interaction in Context Symposium*. New York, NY, USA: ACM, 2014. (IIiX '14), p. 115–124. ISBN 978-1-4503-2976-7. Disponível em: <http://doi.acm.org/10.1145/2637002.2637017>.

MANI, I. *Advances in Automatic Text Summarization*. Cambridge, MA, USA: MIT Press, 1999. ISBN 0262133598.

MANI, I. *Automatic Summarization*. J. Benjamins Publishing Company, 2001. (Natural language processing). ISBN 9789027249869. Disponível em: <https://books.google.com.br/books?id=WVUfl1JsKVQC>.

MANI, I.; BLOEDORN, E.; GATES, B. Using cohesion and coherence models for text summarization. In: *Intelligent Text Summarization Symposium*. [S.l.: s.n.], 1998. p. 69–76.

MANNING, C. D.; RAGHAVAN, P.; SCHüTZE, H. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008. ISBN 0521865719, 9780521865715.

MANNING, C. D.; SURDEANU, M.; BAUER, J.; FINKEL, J.; BETHARD, S. J.; MCCLOSKY, D. The stanford corenlp natural language processing toolkit. In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. [S.l.]: Association for Computational Linguistics, 2014. p. 55–60.

MARCU, D. *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, MA, USA: MIT Press, 2000. ISBN 0262133725.

MARINO, J. B.; E.BANCHS, R.; CREGO, J. M.; GISPERT, A.; LAMBERT, P.; FONOLLOSA, J. A. R.; COSTA-JUSSà, M. R. N-gram-based machine translation. *Comput. Linguist.*, MIT Press, Cambridge, MA, USA, v. 32, n. 4, p. 527–549, dez. 2006. ISSN 0891-2017. Disponível em: <http://dx.doi.org/10.1162/coli.2006.32.4.527>.

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. Disponível em: <http://dl.acm.org/citation.cfm?id=1763653.1763720>.

MCLAUGHLIN, H. G. SMOG grading - a new readability formula. *Journal of Reading*, v. 12, n. 8, p. 639–646, maio 1969.

MCNAMARA, D. S.; GRAESSER, A. C.; MCCARTHY, P. M.; CAI, Z. *Automated Evaluation of Text and Discourse with Coh-Metrix*. New York, NY, USA: Cambridge University Press, 2014. ISBN 0521137292, 9780521137294.

MEENA, Y. K.; DEWALIYA, P.; GOPALANI, D. Optimal features set for extractive automatic text summarization. In: *Proceedings of the 2015 Fifth International Conference on Advanced Computing & Communication Technologies*. Washington, DC, USA: IEEE Computer Society, 2015. (ACCT '15), p. 35–40. ISBN 978-1-4799-8488-6. Disponível em: <http://dx.doi.org/10.1109/ACCT.2015.123>.

MENDOZA, M.; BONILLA, S.; NOGUERA, C.; COBOS, C.; LEóN, E. Extractive single-document summarization based on genetic operators and guided local search. *Expert Syst. Appl.*, v. 41, n. 9, p. 4158–4169, 2014. Disponível em: <http://dblp.uni-trier.de/db/journals/eswa/eswa41.html#MendozaBNCL14>.

MIHALCEA, R.; TARAU, P. Textrank: Bringing order into texts. In: *Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain: [s.n.], 2004. Disponível em: <http://acl.ldc.upenn.edu/acl2004/emnlp/pdf/Mihalcea.pdf>.

MURDOCK, V. G. *Aspects of sentence retrieval*. Tese (Doutorado) — University of Massachusetts Amherst, 2006. AAI3242373.

NANDHINI, K.; BALASUNDARAM, S. R. Use of genetic algorithm for cohesive summary extraction to assist reading difficulties. *Appl. Comp. Intell. Soft Comput.*, Hindawi Publishing Corp., New York, NY, United States, v. 2013, p. 8:8–8:8, jan. 2013. ISSN 1687-9724. Disponível em: <http://dx.doi.org/10.1155/2013/945623>.

NENKOVA, A. *Understanding the Process of Multi-document Summarization: Content Selection, Rewriting and Evaluation*. Tese (Doutorado) — Columbia University, New York, NY, USA, 2006. AAI3203761.

NENKOVA, A.; MCKEOWN, K. Automatic summarization. *Foundations and Trends in Information Retrieval*, v. 5, n. 2-3, p. 103–233, 2011.

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.

NICK, L. *Classifier4J*. 2003. Http://classifier4j.sourceforge.net/. Last acess: March 2015.

O'DONNELL, M. Variable-Length On-Line Document Generation. In: *Proceedings of the 6th European Workshop on Natural Language Generation*. Duisburg, Germany: Gerhard-Mercator University, 1997.

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 Systems with Applications*, v. 65, p. 68 – 86, 2016. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417416304250>.

OLIVEIRA, H.; LIMA, R.; LINS, R. D.; FREITAS, F.; RISS, M.; SIMSKE, S. J. A concept-based integer linear programming approach for single-document summarization. In: *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*. [S.l.: s.n.], 2016. p. 403–408.

OLIVEIRA, H.; LINS, R. D.; LIMA, R.; FREITAS, F. A regression-based approach using integer linear programming for single-document summarization. In: *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*. [S.l.: s.n.], 2017. p. 270–277.

ORăSAN, C. The influence of personal pronouns for automatic summarisation of scientific articles. In: *5th Discourse Anaphora and Anaphor Resolution Colloquium*. Furnas, Portugal: [s.n.], 2004. p. 127–132.

ORăSAN, C. The influence of pronominal anaphora resolution on term-based summarisation. In: NICOLOV, N.; ANGELOVA, G.; MITKOV, R. (Ed.). *Recent Advances in Natural Language Processing V*. Amsterdam & Philadelphia: John Benjamins, 2009, (Current Issues in Linguistic Theory, v. 309). p. 291–300.

OUYANG, Y.; LI, W.; LI, S.; LU, Q. Applying regression models to query-focused multi-document summarization. *Inf. Process. Manage.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 47, n. 2, p. 227–237, mar. 2011. ISSN 0306-4573. Disponível em: <http://dx.doi.org/10.1016/j.ipm.2010.03.005>.

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. *Inf. Process. Manage.*, 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. Disponível em: <http://dl.acm.org/citation.cfm?id=2391258.2391259>.

PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. The pagerank citation ranking: Bringing order to the web. In: *Proceedings of the 7th International World Wide Web Conference*. Brisbane, Australia: [s.n.], 1998. p. 161–172. Disponível em: <citeseer.nj.nec.com/page98pagerank.html>.

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. Disponível em: <http://dx.doi.org/10.1561/1500000011>.

PARDO, T. A. S.; RINO, L. H. M.; NUNES, M. d. G. V. Gistsumm: A summarization tool based on a new extractive method. In: *Proceedings of the 6th International Conference on Computational Processing of the Portuguese Language*. Berlin, Heidelberg: Springer-Verlag, 2003. (PROPOR'03), p. 210–218. ISBN 3-540-40436-8. Disponível em: <http://dl.acm.org/citation.cfm?id=1758748.1758788>.

PARVEEN, D.; RAMSL, H.-M.; STRUBE, M. Topical coherence for graph-based extractive summarization. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2015. p. 1949–1954. Disponível em: <http://www.aclweb.org/anthology/D15-1226>.

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*. AAAI Press, 2015. (IJCAI'15), p. 1298–1304. ISBN 978-1-57735-738-4. Disponível em: <http://dl.acm.org/citation.cfm?id=2832415.2832430>.

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.n.], 2012. p. 2225–2242. Disponível em: <http://aclweb.org/anthology/C/C12/C12-1136.pdf>.

PITLER, E.; NENKOVA, A. Revisiting readability: A unified framework for predicting text quality. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. (EMNLP '08), p. 186–195. Disponível em: <http://dl.acm.org/citation.cfm?id=1613715.1613742>.

PORTER, M. F. An algorithm for suffix stripping. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p. 313–316, 1997.

PRADHAN, S.; RAMSHAW, L.; MARCUS, M.; PALMER, M.; WEISCHEDEL, R.; XUE, N. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*.

Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. (CONLL Shared Task '11), p. 1–27. ISBN 9781937284084.

PRASAD, R. S.; UPLAVIKAR, N. M.; WAKHARE, S. S.; JAIN, V.; YEDKE, T. A. Feature based text summarization. In: *International Journal of Advances in Computing and Information Researches*. [S.l.: s.n.], 2012. v. 1, n. 2.

RADEV, D. R. Experiments in single and multidocument summarization using mead. In: *In First Document Understanding Conference*. [S.l.: s.n.], 2001.

RADEV, D. R.; HOVY, E.; MCKEOWN, K. Introduction to the special issue on summarization. *Comput. Linguist.*, MIT Press, Cambridge, MA, USA, v. 28, n. 4, p. 399–408, dez. 2002. ISSN 0891-2017. Disponível em: <http://dx.doi.org/10.1162/089120102762671927>.

RENNES, E.; JONSSON, A. The impact of cohesion errors in extraction based summaries. In: CHAIR), N. C. C.; CHOUKRI, K.; DECLERCK, T.; LOFTSSON, H.; MAEGAARD, B.; MARIANI, J.; MORENO, A.; ODIJK, J.; PIPERIDIS, S. (Ed.). *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014. ISBN 978-2-9517408-8-4.

RIBALDO, R.; CARDOSO, P. C. F.; PARDO, T. A. S. Exploring the subtopic-based relationship map strategy for multi-document summarization. *RITA*, v. 23, n. 1, p. 183–211, 2016. Disponível em: <http://www.seer.ufrgs.br/index.php/rita/article/view/RITA-VOL23-NR1-183>.

RIEDHAMMER, K.; FAVRE, B.; HAKKANI-TüR, D. Long story short - global unsupervised models for keyphrase based meeting summarization. *Speech Commun.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 52, n. 10, p. 801–815, out. 2010. ISSN 0167-6393. Disponível em: <http://dx.doi.org/10.1016/j.specom.2010.06.002>.

ROTEM, N. *Open Text Summarizer*. 2003. Http://libots.sourceforge.net/. Last acess: Mar. 2015.

SAGGION, H. Creating summarization systems with SUMMA. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*. [s.n.], 2014. p. 4157–4163. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2014/summaries/1102.html>.

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. Disponível em: <http://dl.acm.org/citation.cfm?id=1944566.1944688>.

SAMPSON, G. Briefly noted - english for the computer: The susanne corpus and analytic scheme. *Comput. Linguist.*, MIT Press, Cambridge, MA, USA, v. 28, n. 1, p. 102–103, mar. 2002. ISSN 0891-2017.

SATOSHI, C. N.; SATOSHI, S.; MURATA, M.; UCHIMOTO, K.; UTIYAMA, M.; ISAHARA, H.; INFO-COMMUNICATION, K. H. Sentence extraction system assembling multiple evidence. In: *Proc. 2nd NTCIR Workshop*. [S.l.: s.n.], 2001. p. 319–324.

SCHRIVER, K. A. Evaluating text quality: the continuum from text-focused to reader-focused methods. *IEEE Transactions on Professional Communication*, v. 32, n. 4, p. 238–255, Dec 1989. ISSN 0361-1434.

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.

SILVEIRA, S. B.; BRANCO, A. Enhancing multi-document summaries with sentence simplification. In: *In ICAI 2012: International Conference on Artificial Intelligence, Las Vegas*. [S.l.: s.n.], 2012.

SILVEIRA, S. M. S. B. *Enhancing Extractive Summarization with Automatic Post-processing*. Tese (Doutorado) — University of Lisboa, Alameda da Universidade, 1649-004 Lisboa, Portugal, 2015.

SMITH, C.; HENRIK, D.; ARNE, J. A more cohesive summarizer. In: *COLING (Posters)*. [S.l.: s.n.], 2012. p. 1161–1170.

SMMRY. 2009. Http://smmry.com/. Last acess: Mar. 2015.

SPACHE, G. A new readability formula for primary-grade reading materials. *The Elementary School Journal*, v. 53, n. 7, p. 410–413, 1953. Disponível em: <http://dx.doi.org/10.1086/458513>.

STEINBERGER, J. *Text Summarization within the LSA Framework*. Tese (Doutorado) — University of West Bohemia, 2007. Disponível em: <http://textmining.zcu.cz/publications/PhDThesis-Steinberger.pdf>.

STEINBERGER, J.; JEŽEK, K. Advances in information systems: Third international conference, advis 2004, izmir, turkey, october 20-22, 2004. proceedings. In: _____. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005. cap. Text Summarization and Singular Value Decomposition, p. 245–254. ISBN 978-3-540-30198-1. Disponível em: <http://dx.doi.org/10.1007/978-3-540-30198-1_25>.

STEINBERGER, J.; JEZEK, K. Evaluation measures for text summarization. *Computing and Informatics*, v. 28, n. 2, p. 251–275, 2009. Disponível em: <http://dblp.uni-trier.de/db/journals/cai/cai28.html#SteinbergerJ09>.

STEINBERGER, J.; POESIO, M.; KABADJOV, M. A.; JEEK, K. Two uses of anaphora resolution in summarization. *Inf. Process. Manage.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 43, n. 6, p. 1663–1680, nov. 2007. ISSN 0306-4573. Disponível em: <http://dx.doi.org/10.1016/j.ipm.2007.01.010>.

SUMPLIFY. *Sumplify*. 2015. Http://sumplify.com/. Last acess: Mar. 2015.

SUMY. *Sumy*. 2015. Https://github.com/miso-belica/sumy. Last acess: Mar. 2015.

TAO. *Text Analysis Online ()*. 2015. Http://textanalysisonline.com/simple-text-summarizer. Last acess: Mar. 2015.

TEASER, D. *Py Teaser*. 2013. Https://github.com/xiaoxu193/PyTeaser. Last acess: Mar. 2015.

TEXTCOMPACTOR. 2014. Http://www.textcompactor.com/. Last acess: Mar. 2015.

TEXTSUMMARIZATION. *Text Summarization*. 2015. Http://textsummarization.net/text-summarizer. Last acess: Mar. 2015.

TONELLI, S.; PIANTA, E. Matching documents and summaries using key-concepts. In: *Proceedings of the French Text Mining Evaluation Workshop*. [S.l.: s.n.], 2011.

TOOLS4NOOBS. *Tools4Noobs*. 2015. Http://www.tools4noobs.com/summarize/. Last acess: Mar. 2015.

TORRES-MORENO, J.-M. Automatic text summarization. In: _____. *Automatic Text Summarization*. [S.l.]: John Wiley & Sons, Inc., 2014. p. 23–52. ISBN 9781119004752.

VANDERWENDE, L.; SUZUKI, H.; BROCKETT, C.; NENKOVA, A. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Inf. Process. Manage.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 43, n. 6, p. 1606–1618, nov. 2007. ISSN 0306-4573. Disponível em: <http://dx.doi.org/10.1016/j.ipm.2007.01.023>.

WAN, X.; CAO, Z.; WEI, F.; LI, S.; ZHOU, M. Multi-document summarization via discriminative summary reranking. *CoRR*, abs/1507.02062, 2015.

WANG, D.; LI, T. Weighted consensus multi-document summarization. *Inf. Process. Manage.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 48, n. 3, p. 513–523, may 2012. ISSN 0306-4573. Disponível em: <http://dx.doi.org/10.1016/j.ipm.2011.07.003>.

WANG, S.; LI, W.; WANG, F.; DENG, H. A survey on automatic summarization. In: *Information Technology and Applications (IFITA), 2010 International Forum on*. [S.l.: s.n.], 2010. v. 1, p. 193–196.

WEBSUMMARIZER. 2014. Http://www.websummarizer.com/. Last acess: Mar. 2015.

WHITAKER, R. *NClassifier*. 2004. Http://nclassifier.sourceforge.net/. Last acess: Mar. 2015.

WRITING, C. *Custom Writing Summarizer*. 2006. Http://custom-writing.org/writing-tools/summarizer. Last acess: Mar. 2015.

YEH, J.-Y.; KE, H.-R.; YANG, W.-P.; MENG, I.-H. Text summarization using a trainable summarizer and latent semantic analysis. *Inf. Process. Manage.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 41, n. 1, p. 75–95, jan. 2005. ISSN 0306-4573. Disponível em: <http://dx.doi.org/10.1016/j.ipm.2004.04.003>.

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.

ZAJIC, D. M.; DORR, B. J.; LIN, J. Single-document and multi-document summarization techniques for email threads using sentence compression. *Inf. Process. Manage.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 44, n. 4, p. 1600–1610, jul. 2008. ISSN 0306-4573. Disponível em: <http://dx.doi.org/10.1016/j.ipm.2007.09.007>.