UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE INFORMÁTICA

GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO

TALES TOMAZ ALVES

Busca e recuperação em memórias de tradução utilizando tradução automática e busca semântica

RECIFE

2024

UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE INFORMÁTICA

ENGENHARIA DA COMPUTAÇÃO

TALES TOMAZ ALVES

Busca e recuperação em memórias de tradução utilizando tradução automática e busca semântica

TCC apresentado ao Curso de Engenharia da Computação da Universidade Federal de Pernambuco, como requisito para a obtenção do título de Bacharel em Engenharia da Computação, Centro de Informática da Universidade Federal de Pernambuco.

Orientador(a): Prof. Luciano de Andrade Barbosa

RECIFE

2024

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Alves, Tales Tomaz.
Busca e recuperação em memórias de tradução utilizando tradução automática e busca semântica / Tales Tomaz Alves. - Recife, 2024.
35 p. : il., tab.

Orientador(a): Luciano de Andrade Barbosa
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Informática, Engenharia da Computação - Bacharelado, 2024.
Inclui referências, apêndices.

1. Memória de Tradução. 2. Tradução Automática. 3. Processamento de Linguagem Natural. 4. Representações Vetoriais de Sentenças. 5. Busca Semântica. I. Barbosa, Luciano de Andrade. (Orientação). II. Título.

000 CDD (22.ed.)

TALES TOMAZ ALVES

Busca e recuperação em memórias de tradução utilizando tradução automática e busca semântica

TCC apresentado ao Curso de Engenharia da Computação da Universidade Federal de Pernambuco, como requisito para a obtenção do título de Bacharel em Engenharia da Computação, Centro de Informática da Universidade Federal de Pernambuco.

Aprovado em: 18/03/2024.

BANCA EXAMINADORA

_____

Profº. Dr. Luciano de Andrade Barbosa
Universidade Federal de Pernambuco

_____

Profº. Dr. Kiev Santos da Gama
Universidade Federal de Pernambuco

RESUMO

Tradução é mais do que apenas mudar as palavras de um idioma para outro, ela constrói pontes entre culturas, é um gesto de empatia que permite que pessoas experimentem fenômenos culturais por outra perspectiva. Nesse cenário, entram os sistemas de Memória de Tradução (*Translation Memories*, ou TMs), cuja função central reside na localização e recuperação de traduções existentes em seus bancos de dados para auxiliar tradutores em novas traduções. No entanto, este processo crucial permanece restrito por algoritmos que dependem da distância de edição, o número de remoções, inserções ou substituições necessárias para transformar o segmento de origem no segmento de destino, o que apresenta uma limitação significativa para as TMs modernas. Este trabalho propõe um novo paradigma de sistema de memória de tradução baseado em modelos de tradução automática e representações vetoriais semânticas de sentenças para superar o estado da arte em TMs ao aprimorar a captura de similaridade semântica entre segmentos de texto traduzidos. Os resultados obtidos demonstram que a estratégia é efetiva e promissora, com desempenho significativamente superior a alternativas testadas ($P<0,05$) em traduções de inglês para espanhol e francês, bem como de alemão e português para inglês.

Palavras-chave: Memória de Tradução, Tradução Automática, Processamento de Linguagem Natural, Representações Vetoriais de Sentenças, Busca Semântica.

ABSTRACT

Translation is more than just changing words from one language to another, it builds bridges between cultures, it is a gesture of empathy that allows people to experience cultural phenomena from another culture's lens. In this scenario, enter Translation Memory systems (TMs), whose core function lies in matching and retrieving existing translations in their databases to assist translators in new translations. However, this crucial process remains restricted by algorithms that depend on edit distance, the number of deletions, insertions or substitutions required to transform the source segment into the target segment, which presents a significant limitation for modern TMs. This work introduces an alternative translation memory system paradigm based machine translation models and semantic sentence embeddings to surpass state-of-the-art TMs by improving the capture of semantic similarity between translated text segments. The results obtained demonstrate that the strategy is effective and promising, with significantly superior performance than tested alternatives ($P<0.05$) in translations from English to Spanish and French, as well as from German and Portuguese to English.

Keywords: Translation Memory, Machine Translation, Natural Language Processing, Sentence Embeddings, Semantic Search.

LISTA DE ABREVIAÇÕES

BERT	*Pre-training of Deep Bidirectional Transformers for Language Understanding*
chrF	*Character n-gram F-score*
LaBSE	*Language-agnostic BERT sentence embedding*
NLP	*Natural Language Processing*. Processamento de Linguagem Natural
TA	Tradução Automática
TF-IDF	*Term Frequency - Inverse Document Frequency*
TM	*Translation Memory*. Memória de tradução
USE	*Universal Sentence Encoder*

SUMÁRIO

1 INTRODUÇÃO

Linguagem é muito mais do que uma ferramenta que nos permite comunicar. É a expressão da cultura, da sociedade e das crenças. Há milhares de culturas em todo o mundo, cada uma se expressando em sua própria linguagem no dia a dia, imagine os vastos tesouros culturais perdidos se de repente abandonassem seus dialetos nativos e simplesmente se comunicassem numa língua universal. Outros idiomas podem ser incapazes de expressar certos sentimentos, emoções ou descrições específicas (ŠIMURKA, 2020). Um exemplo disso é a palavra “saudade” que, segundo o dicionário Priberam (2024), significa “lembrança grata de pessoa ausente, de um momento passado, ou de alguma coisa de que alguém se vê privado” uma simples e elegante palavra que expressa um sentimento tão comum a qual não existe em muitas outras línguas como inglês ou francês.

Tradução é mais do que apenas mudar as palavras de um idioma para outro. A tradução constrói pontes entre culturas, é um gesto de empatia que permite que pessoas experimentem fenômenos culturais que, de outra forma, seriam muito estranhos e remotos para serem compreendidos através de suas próprias perspectivas culturais (ŠIMURKA, 2020).

Nesse cenário, um mecanismo que surge para auxiliar os tradutores são as memórias de tradução (TMs) que são bases de dados que armazenam segmentos de texto já traduzidos, a fim de serem reutilizados pelos tradutores em novas traduções. A memória de tradução armazena o texto fonte e sua tradução correspondente em pares de idiomas chamados de “unidades de tradução” (SIMARD, 2020). Os programas de software que usam memórias de tradução são conhecidos como sistemas de memória de tradução.

O conceito de Memórias de Tradução (TMs) criou raízes há mais de quatro décadas. Arthern (1978) observou tradutores da Comissão Europeia desperdiçando um tempo precioso retraduzindo segmentos de texto e propôs um repositório informatizado de textos de fonte e de alvo, facilmente acessível aos tradutores e potencialmente integrado a um sistema computadorizado de terminologia. Com base nesta ideia, muitos sistemas comerciais de TM apareceram no mercado a partir do início da década de 1990.

Desde então, a utilização desta tecnologia específica tem continuado a crescer e estudos demonstram que a utilização de sistemas de memória de tradução

é prevalente entre empresas que produzem documentação multilíngue. Uma pesquisa realizada em 2006 com 874 profissionais de línguas revelou que 82,5% dos participantes utilizavam TMs. A aplicação de TM está associada a textos caracterizados por terminologia técnica e estrutura frasal simples, como documentação técnica em maior grau e, em menor grau, textos de marketing e financeiros (LAGOUDAKI, 2006). Essa correlação se deve à necessidade de habilidades computacionais e à presença de conteúdo repetitivo, características que tornam os sistemas de memória de tradução mais adequados para a tradução de documentação técnica.

Os sistemas de memória de tradução auxiliam os tradutores, sugerindo traduções pré-existentes em seu banco de dados, chamadas correspondências. Essas correspondências são identificadas comparando segmentos para tradução com segmentos armazenados. Existem três categorias de correspondências: correspondências exatas (idênticas aos segmentos armazenados), correspondências difusas (suficientemente semelhantes) e nenhuma correspondência (semelhança insuficiente).

As TMs modernas lançam mão de medidas de similaridade, como distância de edição (LEVENSHTEIN, 1965), ou seja, número de remoções, inserções ou substituições necessárias para transformar o segmento de origem no segmento de destino, para diferenciar correspondências difusas e segmentos não correspondentes. As principais justificativas para o uso desta métrica são o fato de que a distância de edição é um algoritmo simples e rápido de executar em relação à modelos e não é influenciado pela linguagem utilizada. No entanto, naturalmente a distância de edição não consegue capturar a semelhança entre os segmentos quando diferentes palavras e estruturas sintáticas são usadas para expressar a mesma ideia. Como resultado, mesmo que a TM contenha um segmento semanticamente semelhante, o algoritmo de recuperação não será capaz de identificá-lo na maioria dos casos, portanto, a métrica apresenta limitações na captura eficaz da similaridade semântica (GUPTA *et al*., 2016).

Assim, neste trabalho, propõe-se a criação de um novo paradigma de sistema de memórias de tradução baseado em modelos de tradução automática treinados em grandes corpus textuais e em representações vetoriais semânticas de texto com objetivo de expandir o estado da arte e melhorar os resultados já alcançados. Testa-se o sistema em quatro pares de linguagem fonte e alvo: inglês-espanhol,

alemão-inglês, português-inglês e inglês-francês, contudo entende-se que seria possível estender o projeto para quaisquer pares de idiomas e bases de dados.

O trabalho se divide em seis seções como segue. A segunda seção trata de uma revisão de literatura acerca de estratégias para melhorar sistemas de memória de tradução, a terceira de fundamentação teórica em volta de dois temas principais: processamento de linguagem natural e busca semântica, áreas essenciais para o trabalho e que contêm as bases teóricas para construção e avaliação da solução proposta. Na quarta seção, será realizada a formalização da arquitetura e fluxo desenvolvido na solução. Enquanto na seção cinco, serão discutidos os experimentos realizados no trabalho e os resultados atingidos e, por fim, na sexta e última seção serão abordadas as possibilidades de trabalhos futuros e feita a conclusão do trabalho.

2 REVISÃO DE LITERATURA

A busca e recuperação em sistemas de memórias de tradução é um campo em constante evolução, com o objetivo de auxiliar tradutores a encontrar a tradução mais relevante e precisa para um determinado segmento de texto. Atualmente, estes sistemas usam medidas de similaridade simples, como distância de edição calculada nas palavras ou em alguma variação delas (radical, lema), ou seja, não levam em consideração nenhum aspecto semântico na correspondência (GUPTA *et al*., 2016). Na tabela 1 abaixo, pode ser visualizado um exemplo dessa limitação dos algoritmos baseados em distância de edição. No exemplo, a sentença “gato tapete” tem uma distância menor para a referência, embora semanticamente não transmita o mesmo significado como a segunda candidata “No tapete, há um gato”.

Tabela 1 – Exemplo do algoritmo de distância de edição ou de Levenshtein

Frase de referência	Frase candidata	Distância de edição
Há um gato no tapete	gato tapete	**9**
Há um gato no tapete	No tapete, há um gato	17

Fonte: Tales Tomaz Alves (2024).

Nota: Tabela elaborada pelo autor para exemplificar o algoritmo de Levenshtein.

Esforços de pesquisa recentes abordam essa deficiência das TMs baseando-se na arquitetura de redes neurais recorrentes siamesas e no aprendizado de máquina para gerar representações vetoriais das sentenças e obter a similaridade textual semântica, já que obter a similaridade no espaço vetorial é mais intuitivo que comparar textos crus (RANASINGHE; ORASAN; MITKOV, 2019; TAI; SOCHER; MANNING, 2015; MUELLER; THYAGARAJAN, 2016).

Por outro lado, outra área de pesquisa prevalecente é a utilização de *sentence encoders*, ou seja, modelos de NLP geralmente baseados na arquitetura de *Transformers* (VASWANI *et al.*, 2017) que geram representações vetoriais das frases. Nesse cenário, destaca-se a utilização do Universal Sentence Encoder (CER et al., 2018) para capturar segmentos semanticamente semelhantes em TMs melhor do que os métodos baseados na distância de edição (RANASINGHE; ORASAN;

MITKOV, 2020). Em resumo, os autores escolheram o USE devido ao seu desempenho comprovado em tarefas de NLP, às suas capacidades multilingues e ao seu potencial para apoiar os futuros esforços de investigação envolvendo diversos pares de línguas.

Porém, desde então, outros modelos capazes de performar a tarefa de busca semântica multilingue surgiram, com destaque para o BERT (REIMERS; GUREVYCH, 2020), LaBSE (FENG *et al*., 2022) e mE5 (WANG *et al*., 2024), e são investigados mais profundamente neste trabalho.

3 FUNDAMENTAÇÃO TEÓRICA

3.1 Processamento de Linguagem Natural

"Processamento de linguagem natural, NLP (do inglês: *Natural Language Processing*), é o campo que envolve fazer com que os sistemas compreendam as linguagens humanas. Um computador com recursos de NLP seria capaz de receber dados em um idioma como chinês, inglês, hindi, francês, farsi, russo, zulu ou espanhol, e seria capaz de responder da mesma maneira." (GALLAGHER; RAFFERTY; WU, 2004). Esses sistemas são empregados em tarefas diversas, como tradução, reconhecimento ótico de caracteres, reconhecimento de voz, resumo de textos, correção ortográfica entre muitas outras aplicações. A área de pesquisa abrange diversas áreas do conhecimento, sendo elas ciências da computação e da informação, linguística, matemática, engenharia elétrica e eletrônica, inteligência artificial, robótica e psicologia.

Historicamente, diversas abordagens foram utilizadas para se lidar com o problema de processamento de linguagem natural, dentre as quais destacam-se a abordagem simbólica e a abordagem estatística (LIDDY, 2003).

A premissa fundamental da abordagem simbólica pode ser sintetizada através do experimento do quarto chinês idealizado por John Searle (1980). Segundo esta perspectiva, um computador seria capaz de simular a compreensão da linguagem natural (ou outras tarefas de NLP) através da aplicação de um conjunto de regras predefinidas, como um livro de frases em chinês, com perguntas e respostas correspondentes aos dados de entrada.

A partir do final da década de 1980, a introdução de algoritmos de aprendizado de máquina impulsionou uma revolução no campo, impulsionada pelo aumento do poder computacional e pelo declínio das teorias chomskyanas da linguística. Em contraste com as abordagens simbólicas, a abordagem estatística utiliza métodos estatísticos para realizar inferências sobre a distribuição de um conjunto de dados linguísticos, buscando identificar padrões e regularidades sem se basear em regras pré-definidas. Atualmente, essa é a abordagem mais utilizada no processamento de linguagem natural, com destaque para o uso de redes neurais

profundas, que têm apresentado resultados promissores em diversas tarefas, como tradução automática, reconhecimento de fala e análise de sentimento.

3.1.1 Tradução Automática

Tradução automática (TA) é uma área da ciência da computação e da linguística que se dedica ao desenvolvimento de sistemas computacionais capazes de traduzir textos de um idioma para outro. Essa tecnologia tem um papel fundamental na era da globalização, facilitando a comunicação entre pessoas de diferentes culturas, lugares e idiomas.

A área viveu um período de grande evolução nos últimos anos, principalmente impulsionada pela introdução da técnica *Attention* (BAHDANU; CHO; BENGIO, 2014) e da arquitetura *Transformers* em 2017 (VASWANI *et al.*, 2017). Esta arquitetura, baseada em atenção, permite que os modelos de tradução automática concentrem sua atenção em partes específicas da entrada, imitando o foco humano durante a tradução e aprendam relações de longo alcance entre as palavras em uma sentença, resultando em traduções mais precisas e fluentes.

A abordagem utiliza do conceito de *Encoders* e *Decoders*, porém com uma mudança na forma tradicional de modelagem destes. O *Encoder* proposto é responsável por criar uma representação da entrada, de tamanho fixo. Já os *Decoders* são responsáveis por fazer a transformação dos inputs providos pelos Encoders nos resultados finais.

Enquanto a arquitetura *Transformers*, baseada em *Self-Attention*, elimina a necessidade de *Decoders* tradicionais, permitindo que os modelos processem sequências de entrada de forma mais eficiente. Essa arquitetura também facilita o treinamento em grandes conjuntos de dados e a adaptação a tarefas específicas, como tradução automática de documentos ou websites.

Um dos principais avanços nesse campo foi o desenvolvimento do modelo *Transformer-based Neural Machine Translation* (NMT) pela *Google Research*. Esse modelo, treinado em um enorme conjunto de dados de texto traduzido paralelo, obteve resultados significativamente melhores do que os modelos anteriores, estabelecendo um novo estado da arte na TA (WU *et al*., 2016).

Neste trabalho, aplicam-se os modelos de tradução automática especializados e disponibilizados pelo Grupo de Pesquisa em Tecnologia da Linguagem da Universidade de Helsinque, baseados no framework OPUS-MT, concentrado no desenvolvimento de recursos e ferramentas gratuitas para tradução automática (TIEDEMANN; THOTTINGAL, 2020).

3.1.2 Métricas de avaliação

A avaliação da qualidade do texto gerado automaticamente por modelos de NLP é crucial para o desenvolvimento e aprimoramento de tais sistemas. Diversas métricas foram propostas para quantificar a qualidade da geração de texto, focando em diferentes aspectos como fluência, coerência, correção gramatical e fidelidade à referência. Aqui, o foco é a apresentação das duas métricas de avaliação utilizadas: chrF e BLEURT. A primeira, traz uma perspectiva léxica e estrutural para a avaliação, enquanto a última está mais correlacionada com a semântica e a avaliação humana das sentenças. Na tabela 2, visualiza-se essas diferenças de perspectivas entre as métricas. Para a frase referência “Há um gato no tapete” e frase candidata “Um gato está no tapete”, o chrF obteve uma pontuação de 0,57, enquanto o BLEURT de 0,76. Isso indica que a frase candidata apresenta uma adequação semântica maior do que a adequação léxica ou sintática devido às diferenças estruturais entre ela e a referência.

Tabela 2 – Exemplo de resultados das métricas de avaliação selecionadas

Frase de referência	Frase candidata	Métrica	Resultado
Há um gato no tapete	Um gato está no tapete	chrF	0,572
Há um gato no tapete	Um gato está no tapete	BLEURT	**0,756**

Fonte: Tales Tomaz Alves (2024).

Em trabalhos relacionados, utiliza-se BLEU (PAPINENI *et al*., 2002) e METEOR (DENKOWSKI; LAVIE, 2014) para avaliação das traduções, contudo, de acordo com os resultados da competição WMT22 *Metrics Shared Task*, a versão

BLEURT-20 (PU; CHUNG; PARIKH; GEHRMAN; SELLAM, 2021) supera ambas em performance, enquanto o chrF é a melhor métrica léxica não baseada em redes neurais, portanto, estas foram escolhidas.

3.1.2.1 chrF

O chrF (POPOVIĆ, 2015) é uma métrica a nível de sentença que se baseia na comparação de n-gramas de caracteres (uma sequência de n caracteres consecutivos) entre duas sentenças, a de referência e a candidata. A pontuação chrF é calculada como a média ponderada da precisão, *recall* e *F-score* para cada n-grama. Esta métrica pode ser chamada de métrica de sobreposição.

3.1.2.2 BLEURT

BLEURT é uma métrica de avaliação da qualidade de um texto em linguagem natural gerado de maneira automática (SELLAM; DAS; PARIKH, 2020). Ela analisa duas frases, referência e candidata, e pontua sua fluência e fidelidade semântica. Diferente de métricas fixas, BLEURT é um modelo de regressão treinado a partir de notas humanas e, embora vise medir "adequação", correlaciona-se com "fluência" devido a ruído no treinamento. Apesar disso, BLEURT é uma ferramenta automatizada competente para avaliar fluência e significado em textos gerados automaticamente.

Neste projeto, utiliza-se a versão BLEURT-20 (PU; CHUNG; PARIKH; GEHRMAN; SELLAM, 2021) que produz resultados que estão aproximadamente entre 0 e 1 (às vezes menos que 0, às vezes mais que 1).

3.2 Busca semântica

As técnicas tradicionais de banco de dados são excelentes na pesquisa com base em critérios específicos e bem definidos. No entanto, os dados do mundo real muitas vezes exigem a localização de itens semelhantes com base em conceitos vagos ou multifacetados.

Assim, introduz-se o conceito de busca semântica, que tipicamente opera no centro de grandes repositórios de dados, onde os objetos carecem de ordem

inerente e dependem apenas de similaridade semântica para comparação. Por exemplo, ao pesquisar o termo “sapatos pretos” em um site de *e-commerce*, o sistema precisa checar entre bilhões de entradas de produtos. Nesse caso, a busca semântica vai além da correspondência de palavras-chave, considerando variações de terminologia e entendendo o conceito de “sapato preto” em relação a outros produtos (TRIPATHI, 2023).

Isto requer representações semânticas – encapsulamentos que capturam o significado mais profundo dos objetos – para facilitar a identificação eficiente de itens semelhantes.

3.2.1 Representações vetoriais

Representações vetoriais, ou *embeddings*, são o estado da arte quando se trata de representações semânticas textuais, elas convertem palavras e sentenças em vetores numéricos reais, mapeando a similaridade semântica para proximidade no espaço n-dimensional.

A escolha do algoritmo que gera os *embeddings* depende da natureza dos dados e da tarefa. Modelos simples como TF-IDF baseiam-se na frequência de termos, enquanto modelos mais complexos como Word2Vec (MIKOLOV *et al*., 2013), BERT (REIMERS; GUREVYCH, 2020) e USE (CER *et al.*, 2018) utilizam algoritmos de aprendizagem profunda para capturar o contexto e ordem das palavras para gerar suas saídas.

Embeddings de texto abrem um leque de possibilidades para o processamento de linguagem natural, como classificação de texto, análise de sentimento, sistemas de recomendação e, no caso deste trabalho, tradução automática.

3.2.1.1 LaBSE

No trabalho, utiliza-se o modelo BERT gerador de *embeddings* de sentença independente de linguagem (*language-agnostic BERT sentence embedding*, LaBSE) que codifica texto em vetores de alta dimensão (FENG *et al*., 2022). Selecionado por ser treinado e otimizado para produzir representações semelhantes exclusivamente

para pares de frases bilíngues que são traduções uma do outra, o que o torna mais adequado para a extração de traduções de um corpus.

O modelo difere daqueles que geram vetores a nível de palavra porque é treinado em uma série de tarefas de linguagem natural que exigem modelagem do significado, ou seja, semântica, de sequências de palavras em vez de apenas palavras individuais. Portanto, os vetores produzidos também podem ser usados para classificação de texto, agrupamento e outras tarefas de linguagem natural, mas no contexto deste trabalho, é utilizado na busca semântica por produzir representações normalizadas.

3.2.1.2 Multilingual E5

Outro modelo gerador de *embeddings* de sentença utilizado no projeto foi o *Multilingual E5 Text Embeddings* (mE5), que é na verdade um conjunto de modelos de código aberto projetados para resolver as limitações da aplicação de modelos de *embeddings* de texto focados apenas na língua inglesa em tarefas de recuperação da informação e tradução automática. Embora representações vetoriais de texto sejam cruciais para estas tarefas, os modelos existentes muitas vezes carecem de suporte para diversos idiomas, dificultando a sua aplicação no mundo real (WANG *et al.*, 2024).

Os modelos mE5 oferecem três tamanhos (pequeno / base / grande), equilibrando eficiência de inferência e qualidade de incorporação, no contexto do trabalho, utiliza-se a versão base. Eles baseiam-se no sucesso do modelo E5 em inglês, aproveitando um processo de treinamento em duas etapas: pré-treinamento contrastivo não-supervisionado em grande volume de pares de texto, seguido de um ajuste fino supervisionado em pequena quantidade de dados textuais rotulados de alta qualidade.

3.2.2 Distância entre vetores

Na busca semântica, as representações vetoriais permitem a quantificação da similaridade semântica por meio de métricas de distância. Essas métricas medem a proximidade entre vetores no espaço vetorial. A escolha das métricas depende dos

dados subjacentes e dos objetivos da tarefa, mas algumas das métricas de distância comumente usadas são: euclidiana, Manhattan, cosseno e Chebyshev.

3.2.2.1 Similaridade de cosseno

Neste trabalho, utiliza-se a similaridade de cosseno para comparar *embeddings* normalizados de sentenças. Com ela, mede-se o grau do ângulo entre dois vetores, *A* e *B*, e é usada quando a magnitude entre os vetores não importa, apenas a orientação.

Então, dados dois vetores n-dimensionais, *A* e *B*, a similaridade de cosseno, *cos(θ)*, é representada usando um produto escalar e magnitude como:

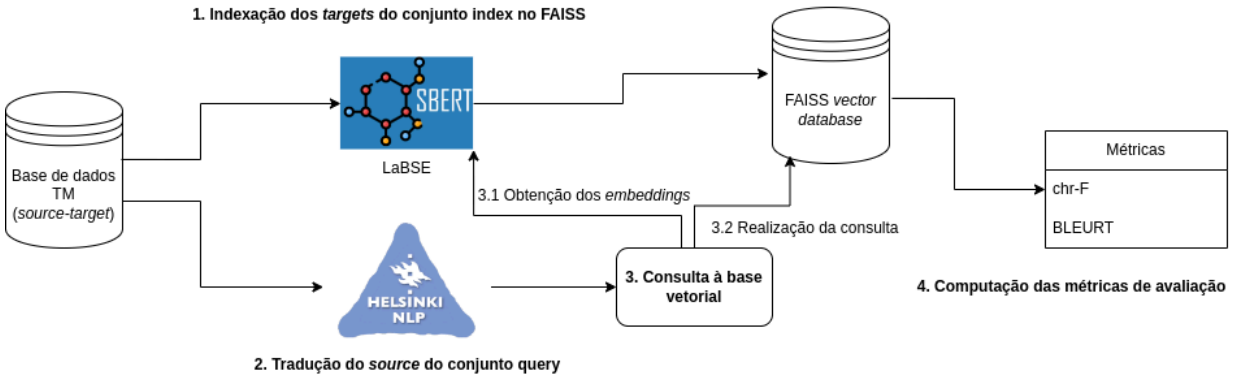
$$cos(\theta) = \frac{A \bullet B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (2.1)$$

Em que A_i e B_i são os i-ésimos componentes dos vetores *A* e *B*, respectivamente. Nesse caso, o valor do cosseno igual a 1 é para vetores que apontam na mesma direção, ou seja, há semelhanças entre as sentenças, em zero para vetores ortogonais, ou seja, não relacionados (mas há alguma semelhança encontrada) e valor -1 para vetores apontando em direções opostas (sem semelhança). Por fim, o termo distância do cosseno é comumente usado para o complemento da similaridade do cosseno no espaço positivo, ou seja, 1 - a similaridade de cosseno.

4 METODOLOGIA

A presente seção descreve a metodologia adotada para investigar e recuperar memórias de tradução utilizando modelos de tradução automática e busca semântica em uma base vetorial. Inicialmente, foram selecionados quatro pares de linguagens *source-target* (alternativamente, fonte-alvo) para traduções: inglês-espanhol (en-es), alemão-inglês (de-en), português-inglês (pt-en) e inglês-francês (en-fr). Para cada par, foi empregado um *dataset* de memórias de tradução distinto, juntamente a um modelo NLP de tradução automática adequado. Dessa forma, para cada par *source-target* no *dataset*, realizou-se o seguinte *pipeline*, detalhado na Figura 1: 1) cálculo de *word embeddings* para as sentenças nas linguagens alvo do conjunto *index* da base de dados escolhida, seguido pelo armazenamento dos vetores gerados em uma base de dados vetorial; 2) tradução das sentenças fonte no conjunto *query* para a respectiva linguagem alvo empregando o modelo selecionado; 3) utilização da tradução obtida, para consultar a base de dados vetorial a partir de *embeddings* gerados a fim de obter o melhor resultado na memória de tradução; 4) computação das métricas de avaliação baseadas no resultado obtido.

Figura 1 – *Pipeline* do projeto



Fonte: Tales Tomaz Alves (2024).

Nota: Imagem elaborada pelo autor para ilustrar o fluxo do sistema implementado na pesquisa.

4.1 Indexação e armazenamento da linguagem alvo

Novamente, para os pares de linguagens inglês-espanhol (en-es), alemão-inglês (de-en), português-inglês (pt-en) e inglês-francês (en-fr), utilizou-se os respectivos *datasets* de memórias de tradução: DGT-TM, uma coleção de memórias de tradução disponível gratuitamente em 22 línguas, disponibilizada pela Direção Geral de Tradução da Comissão Europeia, em conjunto com o Centro Comum de Investigação (STEINBERGER *et al*., 2012); KDE4, um corpus paralelo de arquivos de localização (TIEDEMANN, 2012); *Global Voices*, outro corpus paralelo de notícias do site *Global Voices* compilado e fornecido pela CASMACAT (TIEDEMANN, 2012) e *United Nations*, mais um corpus paralelo composto por registros oficiais e outros documentos parlamentares de domínio público das Nações Unidas (ZIEMSKI; JUNCZYS-DOWMUNT; POULIQUEN, 2016). Em cada base de dados, estão contidos os conjuntos *index*, utilizado para armazenar as sentenças na linguagem alvo no contexto, e o conjunto *query*.

Para armazenar o conjunto *index* das sentenças na linguagem alvo, utilizou-se a base de dados vetorial FAISS-GPU, por sua eficiente busca por similaridade semântica e agrupamento de vetores densos (JOHNSON; DOUZE; JÉGOU, 2019). Para tal, é necessário computar os *word embeddings* normalizados de cada sentença utilizando o modelo multilíngue *Language-agnostic BERT sentence embedding*, LaBSE (FENG *et al.*, 2022).

4.2 Tradução da linguagem fonte

Para cada par de linguagens *source-target*, foi selecionado um modelo de tradução automática apropriado treinado pelo Grupo de Pesquisa em Tecnologia da Linguagem da Universidade de Helsinque com base no *framework* OPUS-MT (TIEDEMANN; THOTTINGAL, 2020). Os modelos utilizados para cada par foram:

- Par inglês-espanhol: Helsinki-NLP/opus-mt-en-es
- Par alemão-inglês: Helsinki-NLP/opus-mt-de-en
- Par português-inglês: Helsinki-NLP/opus-mt-ROMANCE-en
- Par inglês-francês: Helsinki-NLP/opus-mt-en-fr

Cada modelo é empregado na tradução das sentenças na linguagem fonte do conjunto *query* para a respectiva linguagem alvo.

4.3 Consulta à base vetorial

Este passo divide-se em duas etapas. Primeiro, as traduções geradas pelos modelos são convertidas para o mesmo espaço vetorial onde estão armazenadas as sentenças na linguagem alvo do conjunto *index* novamente com o LaBSE (FENG *et al.*, 2022). Na segunda etapa, é realizada uma busca semântica utilizando a métrica cosseno à base vetorial FAISS (JOHNSON; DOUZE; JÉGOU, 2019). Apenas o melhor resultado, ou seja, o vizinho mais próximo, é retornado e rotulado como a tradução da sentença *source*. O resultado é armazenado, juntamente com as respectivas sentenças fonte, alvo e tradução gerada pelo modelo, para análise posterior e obtenção das métricas de avaliação.

4.4 Avaliação

Para analisar os resultados das memórias de tradução recuperadas com base no fluxograma apresentado, compara-se o resultado obtido com a sentença alvo utilizando as seguintes métricas de avaliação:

- chrF (POPOVIĆ, 2015): o *score* a nível de sentença do chrF (*Character n-gram F-score*). No trabalho, utilizou-se por ser a melhor métrica relacionada à estrutura e léxica das sentenças (FREITAG *et al*., 2022).
- BLEURT-20 (PU; CHUNG; PARIKH; GEHRMAN; SELLAM, 2021): métrica de avaliação para geração de linguagem natural. Recebe como entrada o par de sentenças, referência e candidato, e retorna uma pontuação entre 0 e 1 que indica até que ponto o candidato é fluente e transmite o significado da referência. Utilizada pela sua ótima correlação com a avaliação humana das traduções (FREITAG et al., 2022), bem como pelo tamanho do modelo e facilidade de implementação.

5 RESULTADOS

Nessa seção, são apresentados, analisados e contextualizados os resultados do *pipeline* desenvolvido na seção anterior sob a perspectiva das métricas de avaliação chrF e BLEURT. Os resultados foram computados em um ambiente GPU NVIDIA Titan XP com base no conjunto de “teste”, ou *query*, das memórias de tradução para cada par de idiomas. Cada conjunto foi processado após a obtenção das métricas para remover entradas duplicadas ou nulas. A Tabela 3 apresenta a volumetria das diferentes bases utilizadas.

Tabela 3 – Tamanho dos conjuntos *index* e *query* nas bases utilizadas

Base	*Index*	*Query*	*Query* pós-processado
DGT-TM	145490	**49634**	**49632**
Global Voices	**450197**	10000	10000
KDE	160776	10000	9231
United Nations	151480	10000	9545

Fonte: Tales Tomaz Alves (2024).

Nota: Tabela elaborada pelo autor para ilustrar a volumetria das bases de dados.

Para contextualizar os resultados, foi desenvolvido um fluxograma similar ao da seção anterior, utilizando o modelo multilingual-e5-base (WANG *et al*., 2024) para gerar *embeddings* e realizar apenas uma busca semântica multilíngue. O objetivo é comparar o desempenho do *pipeline* completo (tradução + busca semântica) com o da busca semântica pura. Mais especificamente, indexou-se os *embeddings* dos *targets* do conjunto *index* de cada base utilizando o novo modelo; depois, para obter as traduções obtém-se a representação vetorial dos *sources* do conjunto *query* e por fim, realiza-se a busca semântica à respectiva base multilíngue no FAISS.

Na tabela 4, apresenta-se as médias das métricas de avaliação na tradução das respectivas linguagens fonte no conjunto *query* de cada base para o sistema desenvolvido utilizando os respectivos modelos de tradução juntamente à busca

semântica (com o LaBSE) em comparação com os resultados apenas da busca semântica multilíngue (com o multilingual-e5-base).

Tabela 4 – Comparação dos Resultados

Base	Línguas	Técnica	chrF	BLEURT
DGT-TM	inglês-espanhol	Busca semântica (mE5)	0,322	0,336
		Tradução + busca semântica (LaBSE)	**0,354**	**0,353**
Global Voices	português-inglês	Busca semântica (mE5)	0,159	0,078
		Tradução + busca semântica (LaBSE)	**0,253**	**0,367**
KDE	alemão-inglês	Busca semântica (mE5)	0,239	0,388
		Tradução + busca semântica (LaBSE)	**0,307**	**0,463**
United Nations	inglês-francês	Busca semântica (mE5)	0,390	0,311
		Tradução + busca semântica (LaBSE)	**0,391**	**0,324**

Fonte: Tales Tomaz Alves (2024).

Nota: Tabela elaborada pelo autor para mostrar os resultados médios obtidos para cada métrica de avaliação.

Os resultados do *pipeline* foram considerados bons em comparação com a estratégia *baseline* (busca semântica pura), com performance variável entre as bases e idiomas. Na métrica chrF, o melhor resultado foi para a tradução inglês-francês na base *United Nations*, enquanto na métrica BLEURT, o destaque foi para a tradução alemão-inglês na base KDE. O ponto negativo foi a performance da estratégia *baseline* na base *Global Voices*, explicada pela presença de grande quantidade de dados indexados da linguagem alvo (inglês) na verdade escritos na linguagem fonte (português), que acabaram sendo recuperados pelo modelo multilíngue na busca semântica. Na tabela mostrada no Apêndice A, é possível visualizar exemplos de traduções utilizando ambas estratégias.

Para compreender a significância estatística dos resultados obtidos, realizou-se o teste pareado e não-paramétrico de Wilcoxon com nível de significância α em 5% (WILCOXON, 1945), os resultados estão disponíveis no Apêndice B. Através do mesmo, é possível rejeitar a hipótese de que as distribuições são idênticas para as diferentes abordagens e inferir que o sistema construído apresentou desempenho estatisticamente significativamente superior em todos os *datasets* para as métricas selecionadas, BLEURT e chrF. Entretanto, é importante considerar que das bases utilizadas, DGT-TM (2,38%), KDE (0,44%), e *Global Voices* (0,016%) compõem o conjunto de treinamento, OPUS, dos respectivos modelos de tradução utilizados. Apesar disso, a performance semelhante na base *United Nations*, ausente no conjunto de treinamento, demonstra a robustez do projeto.

.

6 CONCLUSÃO

O trabalho investigou a viabilidade de combinar tradução automática com busca semântica para a recuperação de traduções em memórias de tradução. Os resultados demonstram que a estratégia proposta pode ser efetiva, com desempenho estatisticamente significativamente superior ($P<0,05$) à busca semântica pura em todas as bases de dados utilizadas. O sistema desenvolvido apresentou boa performance em diferentes bases e idiomas, com destaque para a tradução inglês-francês na base *United Nations* e alemão-inglês na base KDE.

Vale ressaltar que o desempenho do sistema pode ser aprimorado com a seleção de modelos de tradução e *embeddings* mais adequados para cada par de idiomas e base de dados. A investigação de técnicas para essa seleção é um tópico promissor para pesquisas futuras.

Contudo, o estudo apresenta algumas limitações, como o número relativamente pequeno de *datasets* utilizados, a escolha de pares de linguagens contendo a língua inglesa e a utilização de um único modelo de geração de *embeddings*. Trabalhos futuros podem investigar o desempenho do fluxo em um conjunto mais amplo de bases de dados – sobretudo naquelas não utilizadas no conjunto de treino dos modelos de tradução –, em diferentes tipos de idiomas e com diferentes estratégias e modelos de busca semântica. Além disso, outras áreas de pesquisa promissoras incluem a adaptação do projeto para diferentes tipos de texto, como textos técnicos ou jurídicos e investigação de métodos para avaliar a qualidade das traduções recuperadas de forma mais abrangente, considerando aspectos como fluência, coerência e adequação ao contexto.

De qualquer forma, a combinação de tradução automática com busca semântica multilíngue tem potencial de melhorar significativamente a recuperação de traduções em memórias de tradução. O *pipeline* proposto neste trabalho pode ser aplicado em diversos contextos, como empresas que trabalham com tradução de documentos ou textos, sistemas de tradução automática e ferramentas de tradução assistida por computador (CAT). A pesquisa nesta área pode contribuir para o desenvolvimento de sistemas de tradução automática mais eficientes e precisos, facilitando a comunicação intercultural em diferentes áreas do conhecimento.

Por fim, os resultados da pesquisa demonstram que a combinação de tradução automática com busca semântica é uma estratégia promissora para a

recuperação de traduções em memórias de tradução e abre caminho para futuras pesquisas que possam aprimorar ainda mais a sua efetividade.

REFERÊNCIAS

ARTHERN, Peter J.. *Machine translation and computerized terminology systems: A translator's viewpoint*. *In*: SNELL, Barbara M. (Ed). ***Translating and the Computer***. Londres, Reino Unido: *Aslib Proceedings*, 1978.

BAHDANAU, Dzmitry; CHO, Kyunghyun; BENGIO, Yoshua. *Neural machine translation by jointly learning to align and translate*. *In:* GINSPARG, Paul. **arXiv.** *Cornell University*, 2014. Disponível em: https://arxiv.org/abs/1409.0473. Acesso em: 1 mar. 2024.

CER, Daniel *et al*. *Universal Sentence Encoder*. *In*: RILOFF, Ellen (Ed); CHIANG, David (Ed); HOCKENMAIER, Julia (Ed); TSUJII, Jun'ichi (Ed). ***Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing****: System Demonstrations*. Bruxelas, Bélgica: *Association for Computational Linguistics*, 2018. p. 169–174.

DENKOWSKI, Michael; LAVIE, Alon Lavie. *Meteor universal: Language specific translation evaluation for any target language*. *In*: BOJAR, Ondřej (Ed) *et al*. ***Proceedings of the EACL 2014 Workshop on Statistical Machine Translation***. Baltimore, Estados Unidos da América: *Association for Computational Linguistics*, 2014. p. 376–380.

FENG, Fangxiaoyu; YANG, Yinfei; CER, Daniel; ARIVAZHAGAN, Naveen; WANG, Wei. *Language-agnostic BERT Sentence Embedding*. *In*: MURESAN, Smaranda (Ed); NAKOV, Preslav (Ed); VILLAVICENCIO, Aline (Ed). ***Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)***. Dublin, Irlanda: *Association for Computational Linguistics*, 2022. p. 878–891.

FREITAG, Markus *et al*. *Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust*. *In*: KOEHN, Philipp (Ed) *et al*. ***Proceedings of the Seventh Conference on Machine Translation (WMT)***. Abu Dhabi, Emirados Árabes Unidos: *Association for Computational Linguistics*, 2022. p. 46–68.

GALLAGHER, Shamala; RAFFERTY, Anna; WU, Amy**.** *Natural Language Processing*. *In: Stanford University*. ***NLP - overview*.** Stanford, Estados Unidos da América: *Stanford University*, 2004. Disponível em: https://cs.stanford.edu/people/eroberts/courses/soco/projects/2004-05/nlp/overview.html. Acesso em: 14 fev. 2024.

GUPTA, Rohit; ORASAN, Constantin; ZAMPIERI, Marcos; VELA, Mihaela; VAN GENABITH, Josef; MITKOV, Ruslan. *Improving translation memory matching and retrieval using paraphrases*. ***Machine Translation***, Cham, Suíça, v. 30, p. 19–40, nov. 2016.

JOHNSON, Jeff; DOUZE, Matthijs; JÉGOU, Hervé. *Billion-scale similarity search with GPUs*. ***IEEE Transactions on Big Data***. IEEE, v. 7, n. 3, p. 535–547, jun. 2019.

LAGOUDAKI, Elina. ***Translation Memory systems**: Enlightening users' perspective*. Londres, Reino Unido: *Imperial College London*, 2006.

LEVENSHTEIN, Vladimir Iosifovich. *Binary codes capable of correcting deletions, insertions, and reversals*. ***Proceedings of the USSR Academy of Sciences***. *USSR Academy of Sciences*, v. 163, n.4, p. 845–848, fev. 1965.

LIDDY, Elizabeth D. 2001. *Natural Language Processing*. *In*: ***Encyclopedia of Library and Information Science***. Nova Iorque, Estados Unidos da América: *Marcel Decker, Inc.*, 2001.

MIKOLOV, Tomas; CHEN, Kai, CORRADO, Greg; DEAN, Jeffrey. *Efficient Estimation of Word Representations in Vector Space*. *In*: *International Conference on Learning Representations*, 1., 2013. Scottsdale, Estados Unidos da América. **Anais eletrônicos** [...] Scottsdale: *International Conference on Learning Representations*, 2013. Disponível em: https://iclr.cc/archive/2013/workshop-proceedings.html. Acesso em: 14 fev. 2024.

MUELLER, Jonas; THYAGARAJAN, Aditya. *Siamese Recurrent Architectures for Learning Sentence Similarity*. ***Proceedings of the AAAI Conference on Artificial Intelligence***, Palo Alto, Estado Unidos da América, vol. 30, n. 1, p. 2786–2792, fev. 2016.

PAPINENI, Kishore; ROUKOS, Salim; WARD, Todd; ZHU, Wei-Jing. *BLEU: a Method for Automatic Evaluation of Machine Translation*. *In*: PIERRE, Isabelle (Ed); CHARNIAK, Eugene (Ed); LIN, Dekang (Ed). ***Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics***. Filadélfia, Estados Unidos da América: *Association for Computational Linguistics*, 2002. p. 311–318.

POPOVIĆ, Maja. *chrF: character n-gram F-score for automatic MT evaluation*. *In*: BOJAR, Ondřej (Ed) *et al.* ***Proceedings of the Tenth Workshop on Statistical Machine Translation***. Lisboa, Portugal: *Association for Computational Linguistics*, 2015. p. 392–395.

PU, Amy; CHUNG, Hyung Won; PARIKH, Ankur P.; GEHRMANN, Sebastian; SELLAM, Thibault. *Learning compact metrics for MT*. *In:* MOENS, Marie-Francine (Ed); HUANG, Xuanjing (Ed); SPECIA, Lucia (Ed); YIH, Scott Wen-Tau (Ed). ***Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing***. Punta Cana, República Dominicana: *Association for Computational Linguistics*, 2021. p. 751–762.

RANASINGHE, Tharindu; ORASAN, Constantin; MITKOV, Ruslan. *Intelligent Translation Memory Matching and Retrieval with Sentence Encoders*. *In*: MARTINS, André (Ed) *et al*. ***Proceedings of the 22nd Annual Conference of the European Association for Machine Translation***. Lisboa, Portugal: *European Association for Machine Translation*, 2020. p. 175–184.

RANASINGHE, Tharindu; ORASAN, Constantin; MITKOV, Ruslan. *Semantic textual similarity with Siamese neural networks*. *In*: MITKOV, Ruslan (Ed); ANGELOVA,

Galia (Ed). ***Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)***. Varna, Bulgária: INCOMA Ltd., 2019. p. 1004–1011.

REIMERS, Nils; GUREVYCH, Iryna. *Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation*. *In*: WEBBER, Bonnie (Ed.) *et al*. ***Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)***. Online: *Association for Computational Linguistics*, 2020. p. 4512–4525.

SAUDADE. *In*: Dicionário Priberam da Língua Portuguesa. **Dicionário Online Priberam de Português**. Lisboa, Portugal: Priberam Informática, 2024. Disponível em: https://dicionario.priberam.org/saudade. Acesso em: 11 mar. 2024.

SEARLE, John R. *Minds, brains, and programs*. ***Behavioral and Brain Sciences:*** *An International Journal of Current Research and Theory with Open Peer Commentary*. Cambridge, Reino Unido, v. 3, n. 3, p. 417–424, sep. 1980.

SELLAM, Thibault; DAS, Dipanjan; PARIKH, Ankur. *BLEURT: Learning Robust Metrics for Text Generation*. *In*: JURAFSKY, Dan (Ed) *et al*. ***Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics***. *Online*: *Association for Computational Linguistics*, 2020. p. 7881–7892.

SIMARD, Michel. *Building and using parallel text for translation*. *In*: O'HAGAN, Minako. ***The Routledge Handbook of Translation and Technology***. Londres, Reino Unido: *Routledge*, 2019. cap. 5.

ŠIMURKA, Michal. *The importance of translation*. *In*: Lexika*.* ***Lexika translation and interpreting***. Bratislava, Eslováquia: LEXIKA s.r.o, 2020. Disponível em: https://www.lexika-translations.com/blog/the-importance-of-translation/. Acesso em: 11 mar. 2024.

STEINBERGER, Ralf; EISELE, Andreas; KLOCEK, Szymon; PILOS, Spyridon; SCHLÜTER, Patrick. *DGT-TM: A freely Available Translation Memory in 22 Languages*. *In*: CALZOLARI, Nicoletta (Ed) *et al*. ***Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)***. Istambul, Turquia: *European Language Resources Association (ELRA)*, 2012. p. 454–459.

TAI, Kai Sheng; SOCHER, Richard; MANNING, Christopher D. *Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks*. *In*: ZONG, Chengqing (Ed); STRUBE, Michael (Ed). ***Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)***. Pequim, China: *Association for Computational Linguistics*, 2015. p. 1556–1566.

TIEDEMANN, Jörg. *Parallel Data, Tools and Interfaces in OPUS*. *In*: CALZOLARI, Nicoletta (Ed) *et al*. ***Proceedings of the Eighth International Conference on***

Language Resources and Evaluation (LREC'12). Istambul, Turquia: *European Language Resources Association (ELRA)*, 2012. p. 2214–2218.

TIEDEMANN, Jörg; THOTTINGAL, Santhosh. *OPUS-MT: Building open translation services for the World*. *In*: MARTINS, André (Ed) *et al*. ***Proceedings of the 22nd Annual Conference of the European Association for Machine Translation***. Lisboa, Portugal: *European Association for Machine Translation*, 2020. p. 479–480.

TRIPATHI, Rajat. *What Is Similarity Search?*. *In*: *Pinecone Systems, Inc*. ***Pinecone***. São Francisco, Estados Unidos da América, 2023. Disponível em: https://www.pinecone.io/learn/what-is-similarity-search/. Acesso em: 14 fev. 2024.

VASWANI *et al*. *Attention is all you need*. *In*: VON LUXBURG, Ulrike (Ed) *et al*. ***NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems***. Red Hook, Estados Unidos da América: *Curran Associates Inc.*, 2017. p. 6000–6010.

WANG, Liang; YANG, Nan; HUANG, Xiaolong; YANG, Linjun; MAJUMDER, Rangan; WEI, Furu. *Multilingual E5 Text Embeddings: A Technical Report*. *In:* GINSPARG, Paul. **arXiv.** *Cornell University*, 2024. Disponível em: https://arxiv.org/abs/2402.05672. Acesso em: 29 fev. 2024.

WILCOXON, Frank. *Individual Comparisons by Ranking Methods*. ***Biometrics Bulletin***, Indianápolis, Estados Unidos da América, v. 1, n. 6, p. 80–83, dez. 1945.

WU, Yonghui *et al*. *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. *In:* GINSPARG, Paul. **arXiv.** *Cornell University*, 2016. Disponível em: https://arxiv.org/abs/1609.08144. Acesso em: 1 mar. 2024.

ZIEMSKI, Michał; JUNCZYS-DOWMUNT, Marcin; POULIQUEN, Bruno. *The United Nations Parallel Corpus v1.0*. *In*: CALZOLARI, Nicoletta (Ed) *et al*. ***Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)***. Portorož, Eslovênia: *European Language Resources Association (ELRA)*, 2016. p. 3530–3534.

APÊNDICE A – **Tabela de exemplos de resultados de tradução**

Base (línguas)	Segmento	Texto
DGT-TM (en-es)	Fonte	(ggg) paragraph 2.186 is replaced by the following:
	Alvo	ggg) el punto 2.186 se sustituye por el texto siguiente:
	Resultado tradução + busca	el punto 2.2 se sustituye por el texto siguiente:
	Resultado busca	en el punto 2.2.b), el párrafo segundo se sustituye por el texto siguiente:
Global Voices (pt-en)	Fonte	Célia Possér, Presidente da Plataforma para Direitos Humanos e Equidade de Género reagiu em entrevista a RFI, condenando o acto:
	Alvo	In an interview with RFI, Célia Possér, President of the Platform for Human Rights and Gender Equality, condemned the apparent act of police brutality.
	Resultado tradução + busca	The spokesperson for the UN High Commission for Human Rights, Cecile Pouilly, stated:
	Resultado busca	A principal vítima desses PUAs são pré-adolescentes que atingiram a idade de consentimento, 14 anos.
KDE (de-en)	Fonte	Binden des Ports fehlgeschlagen
	Alvo	Port binding failed
	Resultado tradução + busca	Could not open a socket
	Resultado busca	Port:
United Nations (en-fr)	Fonte	11.12 Together with the World Health Organization (WHO) and the World Meteorological Organization (WMO), UNEP undertook a study on climate change and human health, which was presented to the second Conference of the Parties to the United Nations Framework Convention on Climate Change in July 1996.

	Alvo	11.12 En collaboration avec l'Organisation mondiale de la santé (OMS) et l'Organisation météorologique mondiale (OMM), le PNUE a effectué une étude sur les changements climatiques et la santé, qui a été présentée à la deuxième Conférence des Parties à la Convention-cadre des Nations Unies sur les changements climatiques, en juillet 1996.
	Resultado tradução + busca	45. En coopération avec divers organes et organismes des Nations Unies, le HCDH/CPDH a pris une part active à la préparation et à la célébration de la quatrième Conférence mondiale des Nations Unies sur les femmes, tenue à Beijing en septembre 1995, et de la deuxième Conférence des Nations Unies sur les établissements humains (Habitat II), tenue à Istanbul en juin 1996.
	Resultado busca	h) A pris note d'un projet élaboré par l'OMS, l'OMM et le PNUE de créer un réseau interinstitutions sur les changements climatiques et la santé en vue d'aider les pays à évaluer les incidences des changements climatiques sur la santé et d'améliorer l'accès à l'information pertinente;

Fonte: Tales Tomaz Alves (2024).

Nota: Tabela elaborada pelo autor para exemplificar os resultados obtidos na pesquisa.

APÊNDICE B – **Resultados do teste de Wilcoxon**

Base (línguas)	Resultado - chrF	Valor-P - chrF	Resultado - BLEURT	Valor-P - BLEURT
DGT-TM (en-es)	272 344 784	0	343 050 987	0
Global Voices (pt-en)	4 178 635	0	427 518,5	0
KDE (de-en)	7 271 037	0	6 859 780,5	0
United Nations (en-fr)	13 662 972	0,0005	12 554 564	0

Fonte: Tales Tomaz Alves (2024).

Nota: Tabela elaborada pelo autor demonstrando os resultados do teste estatístico de Wilcoxon.