Uma Investigação sobre Técnicas de Data Augmentation Aplicadas a Tradução Automática Português-LIBRAS

Marcos André Bezerra da Silva

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

João Pessoa, 2024

Marcos André Bezerra da Silva

Uma Investigação sobre Técnicas de Data Augmentation Aplicadas a Tradução Automática Português-LIBRAS

Artigo apresentado ao curso de Ciência da Computação do Centro de Informática, da Universidade Federal da Paraíba, como requisito para a obtenção do grau de Bacharel em Ciência da Computação

Orientador: Tiago Maritan Ugulino de Araújo

Maio de 2024

Catalogação na publicação
Seção de Catalogação e Classificação

S586i Silva, Marcos Andre Bezerra da.
Uma investigação sobre técnicas de data augmentation aplicadas a tradução automática português-LIBRAS / Marcos Andre Bezerra da Silva. - João Pessoa, 2024.
27 f. : il.

Orientação: Tiago Maritan Ugulino de Araújo.
TCC (Graduação) - UFPB/CI.

1. Tradução automática neural. 2. Data augmentation. 3. Libras. I. Araújo, Tiago Maritan Ugulino de. II. Título.

UFPB/CI CDU 004.43

Elaborado por Michelle de Kássia Fonseca Barbosa - CRB-738

CENTRO DE INFORMÁTICA

UNIVERSIDADE FEDERAL DA PARAÍBA

Trabalho de Conclusão de Curso de Ciência da Computação intitulado **Uma Investigação sobre Técnicas de Data Augmentation Aplicadas a Tradução Automática Português-LIBRAS** de autoria de Marcos André Bezerra da Silva, aprovada pela banca examinadora constituída pelos seguintes professores:

Profº. Drº. Tiago Maritan Ugulino de Araújo

Centro de Informática/UFPB

Profº. Drº. Rostand Edson Oliveira Costa

Centro de Informática/UFPB

Profº. Drº. Daniel Faustino Lacerda De Souza

Departamento de Ciências Exatas/UFPB

João Pessoa, 10 de maio de 2024

Centro de Informática, Universidade Federal da Paraıba

Rua dos Escoteiros, Mangabeira VII, João Pessoa, Paraíba, Brasil CEP: 58058-600

Fone: +55 (83) 3216 7093 / Fax: +55 (83) 3216 7117

Agradecimentos

Agradeço à minha mãe Ana por sempre ter me incentivado a seguir na graduação.

Agradeço ao meu orientador e professor Tiago Maritan por sua confiança, apoio e orientações essenciais para o desenvolvimento deste trabalho.

Agradeço também ao Prof. Daniel Faustino, ao Diego Ramon e todos os membros do LAVID que sempre me ajudaram e tornaram este trabalho possível.

Resumo

A tradução automática de Português para LIBRAS é de suma importância para acessibilidade e inclusão de pessoas surdas na sociedade, porém a escassez de dados e o alto custo para construção de um corpus de sentenças autêntico são desafios significativos. *Data augmentation* em Tradução Automática Neural é o processo de geração de sentenças sintéticas a fim de aumentar a quantidade e diversidade do conjunto de treinamento. Este trabalho investiga o uso de técnicas de data *augmentation* para melhoria do desempenho da tradução automática Português-LIBRAS pela métrica BLEU. Dentre as técnicas analisadas, o *back-translation* e sua combinação com substituição por sinônimos com uso de *part-of-speech tagging* se destacaram como as mais eficazes na melhoria do modelo de tradução e podem ser utilizadas para aumentar a diversidade de conjuntos sub-representados no corpus.

Palavras-chave: Tradução Automática Neural, Data Augmentation, Libras

Abstract

The automatic translation from Portuguese to LIBRAS is extremely important for accessibility and inclusion of deaf individuals in society, but the scarcity of data and the high cost of building an authentic corpora pose significant challenges. Data Augmentation in Neural Machine Translation is the process of generating synthetic sentences to increase the quantity and diversity of the training set. This work investigates the use of data augmentation techniques to improve the performance of Portuguese-LIBRAS automatic translation using the BLEU metric. Among the techniques analyzed, *back-translation* and its combination with synonym substitution using *part-of-speech tagging* stood out as the most effective in enhancing the translation model and can be used to increase the diversity of underrepresented datasets.

Key-words: Neural Machine Translation, Data Augmentation, Libras

Sumário

1. Introdução

A evolução da tecnologia tem desempenhado um papel crucial na acessibilidade e inclusão de pessoas surdas, representando um marco significativo na promoção da igualdade de acesso à informação. Uma parcela considerável da população surda possui uma compreensão muito melhor da língua de sinais do que do texto em português. Especialmente na *web*, onde o volume e o dinamismo de informações são enormes, com conteúdo textual sendo gerado a todo instante, a tarefa de interpretar manualmente textos de páginas *web* para língua de sinais é inviável. Diante desse cenário, torna-se indispensável o uso de componentes de tradução automática para traduzir o conteúdo em português para a Língua Brasileira de Sinais (LIBRAS), para que pessoas surdas tenham acesso efetivo à informação online (VERÍSSIMO *et al.*, 2019). Nesse contexto, a plataforma VLibras (ARAÙJO, 2012) emerge como uma solução que torna a *web* verdadeiramente acessível para surdos. O VLibras se destaca pela sua ampla adoção em sites governamentais, onde desempenha um papel essencial ao prover acessibilidade em LIBRAS para os serviços públicos. Sendo utilizado em mais de 500.000 sites públicos e privados, o VLibras realiza milhões de traduções mensalmente (COSTA *et al.*, 2024).

A fim de desenvolver um componente de Tradução Automática entre o português, que é uma língua oral, e a LIBRAS, que é uma língua sinalizada, o componente de tradução do VLibras foi inicialmente desenvolvido utilizando regras de tradução advindas da expertise de linguistas. Na abordagem atual do VLibras, é utilizado um tradutor baseado em Tradução Automática Neural, que é capaz de inferir as regras da linguagem pelos dados de treinamento. A adição do componente baseado em rede neural possibilitou uma maior capacidade de desambiguação de palavras em português que possuem sinais diferentes na LIBRAS, a depender do contexto (VERÍSSIMO *et al.*, 2019).

Entretanto, apesar dos avanços trazidos pelo uso de redes neurais para Tradução Automática, é exigida uma alta quantidade de dados para o treinamento de modelos que gerem traduções de boa qualidade (PONCELAS *et al.*, 2018), o que é um grande obstáculo, principalmente em línguas sinalizadas, como a LIBRAS. É necessário o desenvolvimento de um corpus bilíngue: pares de sentenças em português e suas respectivas traduções em LIBRAS. A produção desse corpus é extremamente custosa, feita manualmente por intérpretes de LIBRAS. É um desafio construir um conjunto de dados de treinamento diverso que representem diferentes contextos de uso da língua (VERÍSSIMO *et al.*, 2019). Neste sentido, a geração de dados sintéticos por técnicas de *data augmentation* é uma estratégia importante para superar a escassez e aumentar a quantidade e diversidade dos dados de treinamento (WANG *et al.*, 2018). Existem vários métodos de *data augmentation* para processamento de linguagem natural e muitos métodos, apesar de poderem ser utilizados em diversas tarefas, foram elaborados com a finalidade de melhorar o desempenho em uma tarefa específica. Até mesmo na tarefa de Tradução Automática, métodos projetados para línguas com muitos recursos disponíveis podem não ser eficazes para línguas com pouco recurso (*low-resource languages*) (FENG *et al.*, 2021). A escolha certa dos métodos empregados para *data augmentation* pode trazer um impacto positivo na qualidade dos modelos gerados. Este trabalho propõe uma investigação sobre o impacto de diferentes

métodos de *data augmentation* no desempenho do modelo de tradução Português-LIBRAS. São exploradas técnicas como *back-translation*, substituição de palavras alinhadas, reversão de *tokens* e substituição por sinônimos com *part-of-speech tagging*, com o objetivo de realizar uma análise quantitativa para comparar o desempenho de diferentes métodos de *data augmentation* e a influência da quantidade de dados sintéticos no desempenho do tradutor através da métrica BLEU4.

2. Fundamentação Teórica

2.1. Tradução Automática Neural

Tradução Automática Neural (*Neural Machine Translation* - NMT) é a aplicação de redes neurais para a tarefa de tradução de sentenças de uma língua de origem para uma língua de destino. Em geral, é utilizada uma rede *sequence-to-sequence*: onde o *encoder* constrói uma representação da sentença no idioma de origem e o *decoder* parte dessa representação e de cada palavra gerada anteriormente pela própria rede para gerar a sentença traduzida no idioma alvo (FADAEE *et al.*, 2017).

Em contraste à Tradução Automática Baseada em Regras (*Rule Based Machine Translation* - RBMT): que transforma a sentença de origem através de algoritmos de substituição por regras, derivadas do conhecimento de linguistas. A Tradução Automática Neural é baseada em dados. É necessária a construção de um corpus bilíngue: um conjunto de pares de sentenças onde, em cada par, uma sentença está na língua de origem e a outra é sua tradução correspondente na língua de destino. As redes neurais treinadas para resolver o problema de tradução são capazes de aprender as regras de tradução diretamente dos dados, eliminando assim a necessidade da construção de algoritmos explícitos com regras de tradução. Especificamente no contexto da LIBRAS, modelos de Tradução Automática Neural oferecem uma capacidade de desambiguação de palavras que têm grafia igual no português, porém significado e sinal diferentes na LIBRAS. Essa capacidade de desambiguação não seria possível de alcançar utilizando apenas uma abordagem de Tradução Automática Baseada em Regras, já que é preciso que o modelo de tradução tenha um entendimento do contexto onde o termo ambíguo está inserido para decidir a desambiguação correta (VERÍSSIMO *et al.*, 2019).

2.2. Data Augmentation

A eficácia da Tradução Automática Neural está intrinsecamente ligada à qualidade e quantidade dos dados utilizados no treinamento dos modelos. As redes neurais revolucionaram o campo da tradução automática, superando abordagens mais tradicionais. No entanto, sua capacidade de fornecer traduções fluentes depende diretamente da disponibilidade de um grande conjunto de exemplos que representem adequadamente o contexto de uso real (PONCELAS *et al.*, 2018).

O principal requisito para um modelo de Tradução Automática de qualidade é a existência de um corpus bilíngue de alta qualidade e extensão significativa. Este corpus serve como base de treinamento para que o modelo de tradução neural aprenda os padrões linguísticos de ambos idiomas de origem e destino (VERÍSSIMO *et al.*, 2019).

Quanto mais abrangente e representativo o corpus, melhor será a capacidade do modelo de generalizar e produzir traduções de qualidade em uma variedade de contextos.

A LIBRAS, assim como outras línguas de sinais, pode ser classificada como língua com poucos recursos (*low-resource language*), tendo em vista a baixíssima quantidade de dados disponíveis para o treinamento de componentes de processamento de linguagem natural (COSTA *et al.*, 2024). A construção do corpus de treinamento é um processo extremamente custoso, demandando esforços intensivos por parte dos linguistas, que criam manualmente cada par de sentenças, resultando em um corpus totalmente elaborado por eles, sem qualquer contribuição de dados pré-existentes.

Para as línguas com poucos recursos, não há dados autênticos traduzidos por humanos suficientes disponíveis para treinar um modelo de Tradução Automática Neural e obter resultados de alta qualidade. A geração de dados sintéticos é uma estratégia interessante para complementar o corpus criado pelos linguistas.

Com um corpus pequeno, o universo de sentenças a serem traduzidas pelo modelo (quando o tradutor estiver sendo utilizado pelo usuário) será muito maior que os exemplos vistos no treinamento. O objetivo de algoritmos de *data augmentation* é, a partir dos dados autênticos, expandir e diversificar o corpus de treinamento com dados sintéticos para que, idealmente, se aproxime da distribuição de dados de todo o universo de pares de sentenças e traduções válidas. A fim de que a quantidade e diversidade desses novos dados beneficiem o modelo de tradução (WANG *et al.*, 2018).

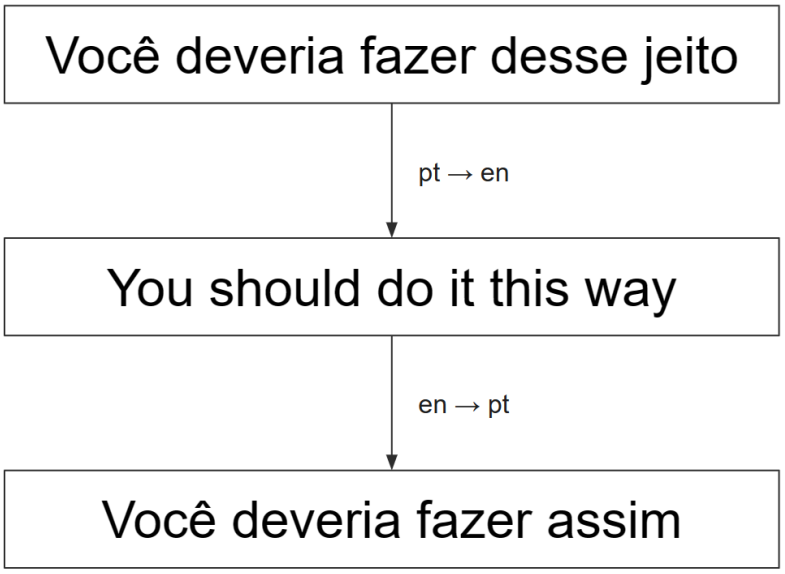
Algoritmos de *data augmentation* geram dados adicionais, sintéticos, a partir dos dados autênticos, através de modificações sobre as sentenças autênticas. Esses dados adicionais são então incorporados ao corpus de treinamento original. Esta é uma tarefa desafiadora no ramo de processamento de linguagem natural e tradução automática, onde qualquer modificação na sentença pode ter impacto no seu significado. Por isso, é importante que as traduções na língua de destino mantenham equivalência com o significado na língua de origem. É uma alternativa mais barata na falta exemplos curados por linguistas e é extremamente importante em línguas com poucos recursos. Porém, por ser um procedimento automático, as sentenças sintéticas geradas pelos algoritmos de *data augmentation* tendem a ser de menor qualidade relativas às sentenças autênticas geradas por humanos. Deve-se, portanto, utilizar uma quantidade razoável de dados sintéticos (PONCELAS *et al*., 2018).

2.2.1. Back-Translation

A técnica de *back-translation* é amplamente utilizada para *data augmentation*, pois tende a gerar sentenças de boa qualidade. Nessa abordagem, é utilizado um outro modelo de tradução e a sentença é traduzida de um idioma para outro e depois de volta para o idioma original, por exemplo, traduzir uma sentença do português para o inglês e, em seguida, de volta para o português. Assim, a partir de um par autêntico de uma sentença na língua de origem e sua respectiva tradução, é possível gerar novas sentenças na língua de origem que tenham o mesmo significado da tradução da sentença original. Isso é feito ao executar o algoritmo nas sentenças do idioma de origem, resultando em novas sentenças sintéticas no mesmo idioma. Essas novas sentenças, resultantes da tradução reversa, são adicionadas ao corpus original no lado do idioma de origem. As traduções correspondentes, na língua de destino, associadas a essas novas sentenças sintéticas serão idênticas às traduções das sentenças autênticas originais, graças a

tendência de preservação do significado (PONCELAS *et al.*, 2018).

Figura 1: Exemplo de *back-translation*



2.2.2. Reversão

A inversão dos *tokens* da sentença na língua de destino é uma tarefa auxiliar e não convencional, porém, faz com que o modelo de tradução utilize mais informações da representação do *encoder* para prever palavras que geralmente aparecem no final da frase, onde a influência do *encoder* tende a diminuir. Dado que as sentenças geradas não são fluentes, é recomendado a adição de um *token* especial no início de cada par de sentenças sintéticas (SÁNCHEZ-CARTAGENA *et al.*, 2021).

2.2.3. Substituição de Palavras

Técnicas baseadas em substituição de palavras envolvem gerar novas sentenças sintéticas escolhendo uma palavra alvo na sentença autêntica para ser substituída aleatoriamente por outra palavra. Isso pode ser aplicado tanto nas sentenças da língua de origem quanto nas da língua de destino, e as substituições podem ser realizadas independentemente entre si. Alternativamente, pode-se respeitar o alinhamento entre as palavras e realizar a substituição aos pares: ao substituir uma palavra na sentença de origem, também é substituída a palavra correspondente na sentença de destino. Mesmo que a substituição seja aleatória, a introdução de ruído nas sentenças de destino ajuda o modelo a aprender a prever a próxima palavra corretamente, mesmo após a geração de uma palavra que não corresponde exatamente ao padrão ouro da tradução humana (WANG *et al.*, 2018; FADAEE *et al.,* 2017).

Substituir palavras por sinônimos é uma abordagem que gera sentenças sintéticas com significado mais próximo da sentença autêntica. Ao escolher uma palavra para ser substituída, é consultada uma tabela de paráfrases que contém sinônimos que são candidatos a substituir a palavra original. Através da representação vetorial da palavra original e das palavras candidatas, é calculada a similaridade cosseno a fim

escolher a palavra candidata com maior similaridade para substituir a palavra original. Também é possível utilizar marcação de partes do discurso *(part-of-speech tagging)* para limitar as opções de substituição dentro da mesma classe gramatical. Ao identificar cada classe gramatical presente em uma sentença por meio de *part-of-speech tagging*, torna-se viável realizar substituições de palavras mantendo a coerência gramatical do texto. Essas restrições também evitam erros que podem ocorrer com o uso do *back-translation*, que confia totalmente na saída do modelo de tradução (MAIMAITI *et al.*, 2021).

Figura 2: Exemplo de reversão e substituição alinhada em tradução alemão-inglês

Task	Lang.	Synthetic training sample
original training sample	source	Es gibt andere Möglichkeiten , die Pyramide zu durchbrechen .
	target	There 's other ways of breaking the pyramid .
reverse	target	. pyramid the breaking of ways other 's There
replace	source	Es gibt aufzurüsten kalt , Schach Spezialwissen zu durchbrechen .
	target	There 's arming cold of breaking chess specialties .

Fonte: SÁNCHEZ-CARTAGENA *et al.* (2021)

Figura 3: Exemplo de substituição de sinônimos com *part-of-speech tagging* para geração de sentenças sintéticas em inglês

Fonte: MAIMAITI *et al.* (2021)

3. Trabalhos Relacionados

FADAEE *et al.* (2017) utilizaram uma estratégia de substituição de palavras alinhadas para aumentar a frequência de palavras raras no conjunto de treinamento. As sentenças geradas foram posteriormente filtradas por um modelo de linguagem que foi treinado com o objetivo de avaliar se as sentenças sintéticas são fluentes, ou seja, estão corretas gramaticalmente e fazem sentido semântico. Foi observado um ganho de 2.5 pontos segundo a métrica BLEU na tradução de inglês para alemão.

Sánchez-Cartagena *et al.* (2021) utilizaram duas tarefas para *data augmentation*. Uma delas é uma substituição de palavras alinhadas semelhante à apresentada por FADAEE et al. (2017), porém sem se preocupar com a fluência das sentenças geradas, não sendo necessário um modelo de linguagem adicional. De maneira auxiliar, também foi utilizada a tarefa de reversão de *tokens.* Foi avaliada a tradução entre inglês e alemão, hebreu e vietnamita resultando em um ganho médio de 1.6 BLEU.

MAIMAITI *et al.* (2021) realizaram substituição de sinônimos com *part-of-speech tagging* para tradução dos idiomas azerbaijão, hindi, uzbeque, turco, alemão e chinês para inglês, observando um ganho de BLEU entre 1.16 e 2.39 pontos.

JANG *et al.* (2022) utilizaram técnicas de *data augmentation* para língua de sinais coreana (*Gloss-level Korean Sign Language*). Utilizaram *back-translation*, substituição por sinônimos restrita às classes de substantivos, nomes próprios e pronomes, além de substituição de palavras utilizando um modelo de linguagem coreano. Resultando em ganhos de BLEU em 10, 12 e 16 pontos, respectivamente.

WANG, YANG (2022) trabalharam em modelos de tradução entre os idiomas inglês, chinês e tailandês. Observaram ganhos de 1 BLEU ao utilizar substituição de palavras alinhadas na tradução de chinês para inglês e 4 de BLEU na tradução de inglês para chinês. Utilizaram substituição por sinônimos na tradução de chinês para tailandês, resultando em um ganho de 2,7 BLEU.

4. Metodologia

4.1. Técnicas Selecionadas

A fim de realizar experimentos com o objetivo de identificar quais métodos de *data augmentation* trazem o melhor ganho de desempenho ao tradutor. Foram selecionadas: (a) *back-translation,* por ser amplamente utilizada em Tradução Automática Neural, (b) a reversão e substituição como apresentados por SÁNCHEZ-CARTAGENA *et al.* (2021), por não necessitar de modelos de linguagem adicionais e ser uma iteração de métodos anteriores e (c) a substituição por sinônimos com *part of speech tagging* como proposto por MAIMAITI *et al.* (2021), por tender a gerar sentenças fluentes.

Para o *back-translation*, são utilizados dois modelos de tradução automática disponíveis no *framework Hugging Face Transformers*: o modelo *Helsinki-NLP/opus-mt-ROMANCE-en*, que traduz do português para o inglês, e o modelo *Helsinki-NLP/opus-mt-en-ROMANCE*, que traduz do inglês de volta para o português. Durante a etapa de *back-translation,* as sentenças em português do corpus autêntico são traduzidas para o inglês pelo primeiro modelo e, em seguida, traduzidas de volta para o português pelo segundo modelo. É importante ressaltar que, em algumas situações, a sentença gerada pela tradução de volta pode ser idêntica à sentença original

em português. Apesar disso, mesmo quando ocorrem duplicatas entre as sentenças originais e as sentenças sintéticas geradas pelo *back-translation*, o processo ainda resulta em um significativo crescimento do corpus de treinamento. Após remover as duplicatas, o corpus expandido apresenta um aumento de cerca de 50% no número total de sentenças.

A substituição de palavras alinhadas (*replace*), como apresentada por SÁNCHEZ-CARTAGENA *et al.* (2021), foi realizada utilizando o modelo de linguagem estatístico MOSES que utiliza a biblioteca GIZA. O MOSES é um sistema de tradução automática que opera através do alinhamento um para um de todos os *tokens* da língua de origem para a língua de destino. Esse modelo de Tradução Automática Estatística produz um léxico que contém entradas de palavras na língua de origem, juntamente com a probabilidade associada de que tal palavra na língua de destino seja uma tradução apropriada. Para a substituição, foi determinado substituir uma palavra por sentença. É sorteado aleatoriamente um par de palavras alinhadas que tenham uma probabilidade maior que 0.7 no léxico. Em seguida, esse par de palavras é substituído por outro par de palavras, também de probabilidade maior que 0.7, proporcionando uma variação semântica na sentença. O método *reverse* também foi implementado, esse método consiste em inverter a lista dos *tokens* da *string* da sentença original, separados pelo caractere de espaço. Para mitigar o impacto de sentenças potencialmente não fluentes, foi adicionado um *token* especial precedendo a sentença de origem em cada método. Essa medida tem como objetivo diminuir a influência dessas sentenças no resultado final do tradutor.

A implementação da substituição por sinônimos baseada em *part-of-speech tagging*, conforme descrito por MAIMAITI *et al*. (2021), utiliza uma combinação de técnicas e recursos de processamento de linguagem natural, incluindo modelos de *part-of-speech tagging*, a base de dados *WordNe*t para identificar sinônimos candidatos e *word embeddings* para calcular a similaridade entre os sinônimos candidatos e escolher o sinônimo com maior similaridade cosseno. Para realizar a marcação de partes do discurso, foi utilizado o modelo *POS_tagger_brill.pkl*, disponível no repositório do *GitHub inoueMashuu/POS-tagger-portuguese-nltk* para o framework de processamento de linguagem natural *nltk*. Esse modelo é responsável por atribuir classes gramaticais às palavras em português para a identificação das palavras alvo que serão substituídas por sinônimos. A base de dados *WordNet*, acessada através da biblioteca *nltk*, foi empregada como fonte de sinônimos para as palavras identificadas pelo *part-of-speech tagging*. O *WordNet* é uma ampla base de dados lexical que organiza palavras em conjuntos de sinônimos, conhecidos como *synsets*, e fornece informações sobre suas relações semânticas e gramaticais. Essa base de dados permite a consulta de sinônimos restritos a classes gramaticais específicas, como sujeitos, adjetivos, advérbios e verbos, permitindo que a substituição gere sentenças de maior qualidade. Por fim, para calcular a similaridade entre as palavras alvo e os sinônimos disponíveis no *WordNet*, foram utilizadas as *embeddings skip-gram* de 100 dimensões do *Repositório de Word Embeddings* do *NILC*. As *word embeddings* são representações vetoriais das palavras que capturam suas relações semânticas com base em seu contexto de ocorrência. Essas representações vetoriais permitem calcular a similaridade cosseno entre palavras, necessário para identificar o sinônimo mais adequado para a substituição.

4.2. Ambiente de Treinamento

O treinamento foi realizado utilizando o *framework fairseq,* desenvolvido pelo *facebook*, projetada com foco no treinamento de redes *sequence-to-sequence*, que são adequadas para tarefas de tradução automática. Dentro do *framework*, optou-se por uma versão reduzida de um modelo que adota a arquitetura *Transformer*, proposta por Vaswani *et al.* (2017). No *fairseq*, o modelo *transformer* reduzido *transformer_iwslt_de_e*n foi pré treinado na tradução de alemão para inglês e tem uma quantidade de cabeças de atenção reduzida. A escolha desse modelo específico para o treinamento de tradução Português-Libras permite que o treinamento seja muito mais rápido. Fazer o ajuste fino (*fine-tuning*) de um modelo de linguagem já treinado, mesmo que em um par de idiomas diferentes, tende a ser mais rápido do que treinar um modelo do zero (RANATHUNGA *et al.*, 2021). Além disso, o fato de o modelo ser reduzido oferece vantagens adicionais. Modelos menores geralmente exigem menos recursos computacionais durante o treinamento e a inferência, o que é interessante quando os recursos de *hardware* são limitados. No caso específico deste trabalho, a disponibilidade de GPUs T4 gratuitas na plataforma *Google Colab* torna o treinamento do modelo mais acessível. Outro aspecto crucial é a consideração do tempo de treinamento. Modelos menores tendem a treinar mais rapidamente do que seus equivalentes maiores, permitindo iterações mais rápidas no processo de desenvolvimento e ajuste do modelo. Além de possibilitar que os treinamentos sejam executados em qualquer conta gratuita do *Google Colab.*

O tradutor do VLibras possui uma arquitetura híbrida de Tradução Automática Baseada em Regras e Tradução Automática Neural. A sentença em português é pré-processada por um componente de tradução baseada em regras, a saída do pré-processamento alimenta o componente de rede neural do tradutor, que é treinado com o objetivo de aproximar a glosa gerada pelo tradutor de regras para a glosa gerada pelos intérpretes humanos (COSTA *et al.,* 2024).

A LIBRAS é uma língua composta por de gestos, expressões faciais e movimentos corporais que carregam significados linguísticos, muito diferente das línguas orais que geralmente possuem um sistema de escrita baseado na representação dos sons. O corpus do VLibras possui pares de sentenças em português como língua de origem e uma representação intermediária de glosas em LIBRAS para as sentenças no idioma de destino. O corpus mencionado consiste em aproximadamente 65.000 exemplos de pares de sentenças (COSTA *et al.*, 2024), cobrindo uma ampla variedade de tópicos e contextos linguísticos. As glosas são criadas por intérpretes e linguistas especializados na língua de sinais e cada sinal da glosa possui uma animação associada. O uso de glosas intermediárias como uma forma de representação linguística da língua de sinais permite que os algoritmos de PLN trabalhem de forma mais eficaz com a LIBRAS (LIMA *et al.*, 2020).

No *pipeline* de tradução do VLibras, já existem cinco métodos de *data augmentation* que geram sentenças fluentes e foram construídos com o conhecimento de linguistas da LIBRAS. O primeiro método, denominado "Lugares", tem como objetivo introduzir variação nas sentenças ao identificar nomes de lugares, como cidades, estados ou países, e substituí-los por outras localidades disponíveis em tabelas de substituição auxiliares. O segundo método, chamado de "Negação", trabalha na identificação de sinais na frase que possam ser negados e gera novas sentenças

realizando essa substituição. Essa técnica é essencial para aprimorar a capacidade do sistema de tradução em compreender e gerar corretamente sentenças negativas, um aspecto importante da gramática da LIBRAS. O terceiro método, denominado "Intensidade", foca na identificação de advérbios na frase que podem ser substituídos para gerar novas frases com diferentes níveis de intensidade. Por exemplo, a substituição de "muito" por "pouco" ou vice-versa. O quarto método, conhecido como "Famosos", visa diversificar o conteúdo das sentenças ao identificar sinais referentes a pessoas famosas e substituí-los por outros sinais representando outros famosos. Essa técnica melhora a capacidade de desambiguação do sistema ao lidar com pessoas conhecidas. Por fim, o quinto método, denominado "Direcionalidade" visa capturar uma nuance gramatical específica da língua de sinais, levando em consideração o sujeito e o objeto do verbo. Na língua de sinais, diferentes sinais são utilizados para representar a mesma palavra em português, dependendo do sujeito e do objeto da mensagem. O método de Direcionalidade identifica esses sinais na sentença e realiza substituições apropriadas com outros sinais direcionais equivalentes. Todas essas técnicas de aumento de dados são aplicadas sobre as sentenças do corpus inicialmente construído por linguistas especializados em LIBRAS. As sentenças sintéticas geradas por essas técnicas são então anexadas ao corpus original, formando o corpus base de treinamento, que conta com aproximadamente 106 mil sentenças.

Durante o treinamento é feita uma divisão, conhecida como *split*, das sentenças autênticas do corpus. A divisão das sentenças segue uma distribuição que reserva 30% das sentenças para validação e utiliza os 70% restantes para o treinamento. Os pares de sentenças reservadas para o treinamento seguem para ser processadas por outros elementos do *pipeline* de treinamento. Essas sentenças são processadas por cada método de *data augmentation* padrão no *pipeline*, formando o conjunto de dados de treinamento que é o *baseline* para os experimentos realizados, com cerca de 106 mil pares de sentenças. Em seguida, cada método de *data augmentation* implementado neste trabalho é aplicado sobre o conjunto de sentenças do *baseline*. Nesta etapa, quando há combinação de diferentes métodos, a entrada de cada método de *data augmentation* é restrito às sentenças do *baseline*. No entanto, uma exceção é feita para o método de *back-translation* devido a restrições de recursos computacionais. Nesse caso, as sentenças sintéticas geradas pelo *back-translation* são anexadas às sentenças autênticas antes da execução dos métodos de *data augmentation* padrão do *pipeline* a fim de otimizar recursos e diminuir o custo com o uso de GPU, tendo em vista que não foi possível executar o *back-translation* gratuitamente no *google colab*. A semente de geração de números aleatórios é fixada em todos os experimentos realizados durante o treinamento do modelo, isso garante a reprodutibilidade dos resultados e permite uma comparação justa entre diferentes treinamentos. Por fim, a performance do modelo é avaliada através de um conjunto de avaliação cuidadosamente selecionado por linguistas, contendo 50 sentenças para cada grupo relevante no domínio da LIBRAS. Esses grupos incluem sentenças básicas, cardinais, de contexto, relacionadas à direcionalidade, a pessoas famosas, à intensidade, a lugares e à negação, permitindo uma avaliação do desempenho do modelo em diferentes contextos linguísticos.

4.3. Métricas de Interesse

Avaliar a qualidade das traduções geradas por modelos de Tradução Automática Neural é uma tarefa desafiadora devido à natureza complexa e subjetiva da linguagem. Diferentes traduções podem ser consideradas aceitáveis para uma mesma sentença de origem, dependendo de uma variedade de fatores como contexto, estilo e preferências individuais. Uma das métricas mais amplamente utilizadas para avaliar a qualidade da tradução automática é o BLEU (*Bilingual Evaluation Understudy*) (PAPINENI *et al.* 2002). O BLEU é uma métrica que varia de 0 a 100 e busca automatizar e replicar como um humano julgaria a qualidade da tradução. Essa métrica é baseada na comparação dos n-gramas da sentença gerada com as traduções de referência disponíveis, a fim de calcular a similaridade entre a tradução gerada pelo modelo e a tradução de referência. O BLEU4, por exemplo, é calculado com base em n-gramas de quatro palavras: Isso significa que o modelo é avaliado com base na precisão dos n-gramas de quatro palavras em suas traduções, em comparação com as traduções de referência disponíveis. Uma das principais vantagens do BLEU é sua rapidez de cálculo, o que o torna uma métrica eficiente para avaliação automática. No entanto, é importante ressaltar que o BLEU apresenta algumas limitações. Por exemplo, ele não leva em conta considerações sobre similaridades semânticas entre as traduções, o que pode resultar em pontuações imprecisas em alguns casos. Neste trabalho, a análise dos resultados é realizada com base no desempenho do modelo de tradução seguindo a métrica BLEU4.

5. Resultados e Discussões

Foram realizadas duas rodadas de experimentos com o objetivo de analisar o impacto de diferentes métodos de *data augmentation* no desempenho do modelo de tradução. Na primeira rodada de experimentos, cada método de *data augmentation* foi aplicado em sequência sobre o *baseline* sem restrição sobre a quantidade de sentenças sintéticas geradas. Posteriormente, na segunda rodada de experimentos, o tamanho do conjunto de treinamento foi limitado a 155 mil sentenças, buscando um equilíbrio entre a quantidade de sentenças autênticas e de sentenças sintéticas.

Todos os métodos de *data augmentation* apresentaram melhorias sobre o *baseline*, conforme observado na Tabela 1. O *back-translation* trouxe um ganho significativo (+15,82 BLEU) mesmo com o menor acréscimo de dados (51 mil sentenças). Tanto a reversão com substituição de palavras alinhadas (*reverse + replace*), quanto a substituição baseada em *part-of-speech tagging* demonstraram resultados próximos (50,73 e 50,32 pontos de BLEU, respectivamente).

Tabela 1: Resultado em BLEU do desempenho do modelo de tradução utilizando as técnicas de *data augmentation* selecionadas.

	Baseline (106k)	**Back Translation (155k)**	**Reverse + Replace (313k)**	**Part of Speech (346k)**
Básicas	42,05	51,66	47,89	52,67
Cardinais	50,18	64,78	72,86	70,52
Contexto	24,37	43,92	47,67	38,86
Direcionalidade	0,00	18,57	30,28	29,06
Famosos	27,34	45,07	43,23	40,70
Intensidade	38,47	49,38	42,5	36,55
Lugares	44,04	56,68	51,85	55,04
Negação	44,04	50,84	57,14	62,42
Romanos	25,86	57,85	63,15	67,12
Média	32,93	48,75	**50,73**	50,32

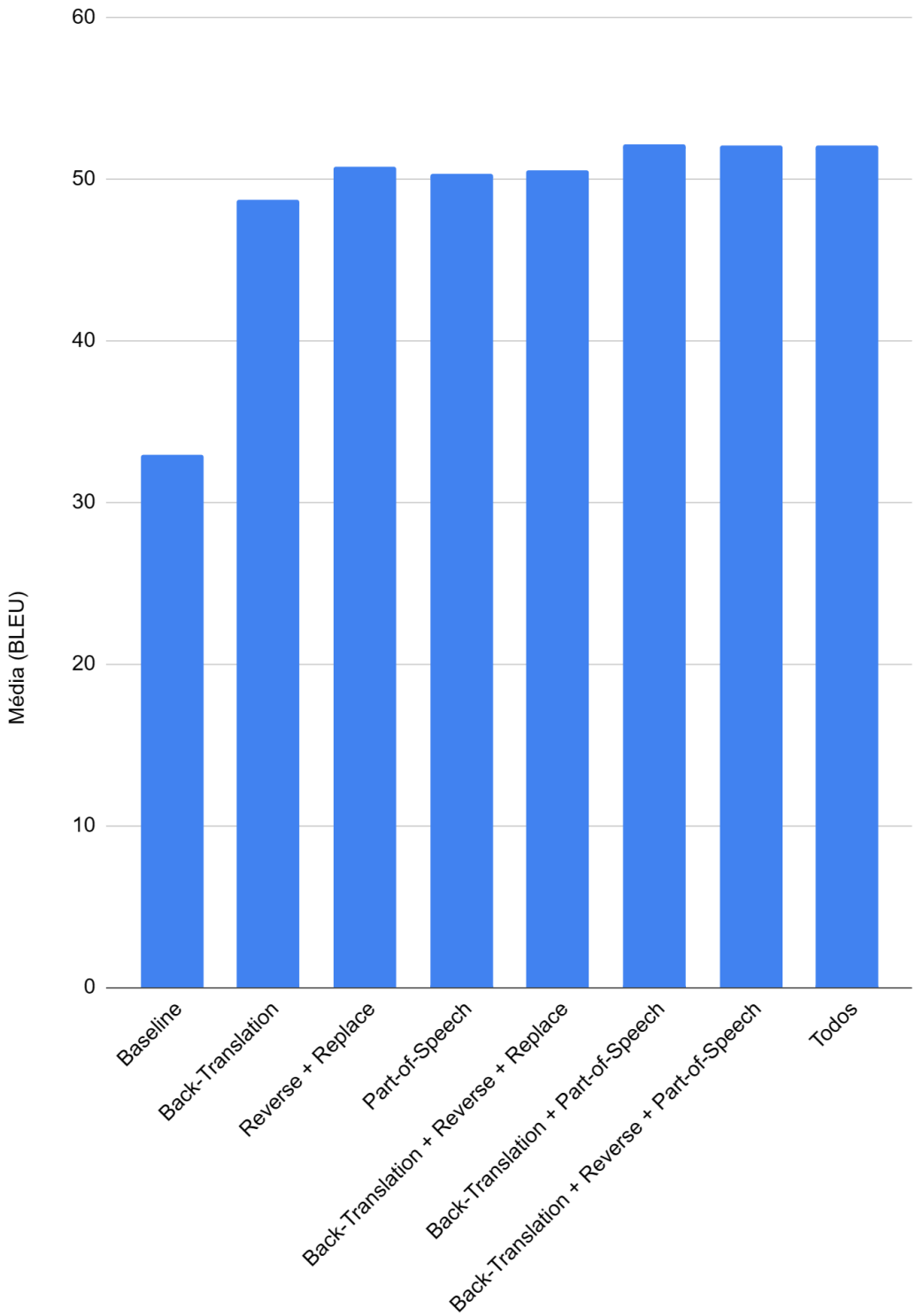
Ao adicionar o *back-translation* em conjunto com a combinação *reverse + replace*, conforme a Tabela 2, foi observada uma pequena piora não significativa no resultado geral. Por outro lado, a combinação do *back-translation* com a substituição baseada em *part-of-speech tagging* resultou em um ganho expressivo de desempenho (+19,25 em relação ao *baseline*), possivelmente devido ao aumento substancial da quantidade de sentenças resultantes dessa combinação. No entanto, ao incluir mais métodos de *data augmentation*, mesmo que isso aumente ainda mais a quantidade de sentenças no conjunto de treinamento, não foi observada uma melhoria significativa no desempenho do modelo. Isso sugere que existe uma limitação na melhoria do desempenho proporcionada pela quantidade de sentenças sintéticas geradas por *data augmentation*, especialmente em função da quantidade de dados autênticos disponíveis. Esses resultados evidenciam a importância de encontrar um equilíbrio adequado entre a

quantidade de dados autênticos e sintéticos no conjunto de treinamento.

Tabela 2: Desempenho em BLEU de combinações de técnicas de *data augmentation* sem restrição na quantidade de sentenças

	Baseline (106k)	**Back Translation + Reverse + Replace (467k)**	**Back Translation + Part of Speech (540k)**	**Back Translation + Reverse + Part of Speech (695k)**	**Todos (851k)**
Básicas	42,05	48,55	48,14	50,76	53,53
Cardinais	50,18	61,54	67,29	66,79	66,49
Contexto	24,37	45,68	43,73	39,36	49,25
Direcionalidade	0,00	27,66	33,60	36,95	32,66
Famosos	27,34	43,74	48,09	44,83	42,85
Intensidade	38,47	45,37	39,61	42,29	42,29
Lugares	44,04	54,45	48,10	52,42	47,12
Negação	44,04	53,84	65,94	59,17	59,34
Romanos	25,86	74,17	75,19	76,07	75,21
Média	32,93	50,55	**52,18**	52,07	52,08

Figura 4: Média em BLEU do desempenho das técnicas de *data augmentation* sem restrição no tamanho do conjunto de treinamento.



Na segunda fase dos experimentos, onde o tamanho do conjunto de treinamento foi limitado, o método de *back-translation* apresentou os melhores resultados (+15,82 BLEU sobre o *baseline*) no cenário apresentado pela Tabela 3, onde cada método foi

testado isoladamente. Esse resultado não é surpreendente, uma vez que o *back-translation* é um dos métodos mais estabelecidos e amplamente utilizados para *data augmentation* em processamento de linguagem natural.

Tabela 3: Resultado em BLEU de cada técnica utilizada isoladamente, com restrição de 155 mil sentenças no conjunto de treinamento

	Baseline	**Back Translation**	**Part of Speech**	**Reverse**	**Replace**
Básicas	42,05	51,66	48,74	43,07	48,51
Cardinais	50,18	64,78	56,29	60,33	60,69
Contexto	24,37	43,92	35,00	32,80	37,85
Direcionalidade	0,00	18,57	0	0	22,34
Famosos	27,34	45,07	47,41	34,36	45,23
Intensidade	38,47	49,38	38,54	38,65	43,14
Lugares	44,04	56,68	51,5	47,87	47,73
Negação	44,04	50,84	58,87	42,95	53,6
Romanos	25,86	57,85	40,72	31,07	57,25
Média	32,93	**48,75**	41,89	36,78	46,26

Ao considerar um cenário com aumento de 25% nos dados para cada método empregado, foram testadas diversas combinações de métodos aos pares, conforme visto na Tabela 4. Notavelmente, a combinação de *back-translation* com *part-of-speech* demonstrou o melhor desempenho quando há a limitação do tamanho do conjunto de treinamento, com um ganho de 16,5 BLEU sobre o *baseline*. Essa combinação já havia apresentado bons resultados quando não há limitação no tamanho do conjunto de treinamento, com 19,25 pontos de BLEU sobre o *baseline,* isso mostra que as duas técnicas produzem resultados bons quando usadas em conjunto. Uma possível explicação para o bom desempenho dessa combinação é o aumento da diversidade do

vocabulário proporcionado pelo *back-translation*. Ao traduzir as sentenças de volta para o idioma original, o conjunto de treinamento conta com um vocabulário mais amplo, o que por sua vez amplia as opções de substituição por sinônimos para o *part-of-speech*. Devido à restrição na quantidade de sentenças, a eficácia do método *reverse* acabou sendo reduzida quando combinado com outras técnicas. Essa é uma desvantagem que o experimento traz para esse método, porque ele é sugerido como uma tarefa destinada a fortalecer o *encoder* de forma independente de outras técnicas de *data augmentation*. Além das combinações de pares de métodos, também foram exploradas outras configurações envolvendo três técnicas distintas.

Tabela 4: Resultado em BLEU do treinamento do modelo de tradução com contribuição de 25 mil sentenças por método

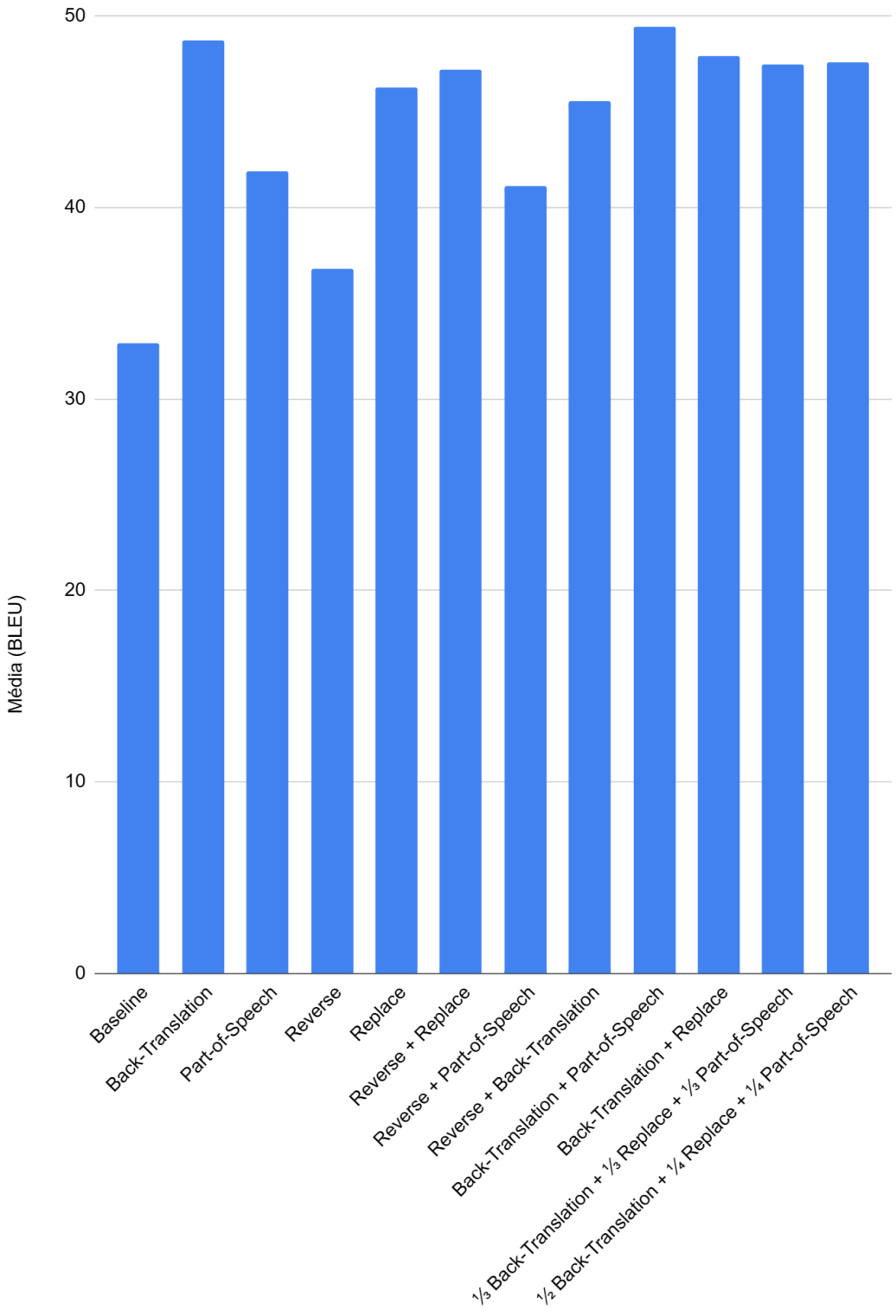
	Baseline	**Reverse + Replace**	**Reverse + Part of Speech**	**Back Translation + Part of Speech**	**Back Translation + Reverse**	**Back Translation + Replace**
Básicas	42,05	47,29	40,84	48,37	47,86	46,25
Cardinais	50,18	66,69	57,17	66,33	67,18	64,80
Contexto	24,37	43,86	33,08	38,86	38,56	43,28
Direcionalidade	0,00	18,61	17,12	28,12	21,91	25,01
Famosos	27,34	52,12	41,58	46,67	42,07	46,46
Intensidade	38,47	40,21	49,63	43,62	44,40	42,06
Lugares	44,04	48,65	47,97	50,84	52,17	52,65
Negação	44,04	57,69	48,55	60,05	46,69	50,59
Romanos	25,86	49,51	34,07	62,09	48,99	60,08
Média	32,93	47,18	41,11	**49,43**	45,53	47,90

Foram testadas mais algumas configurações, conforme a Tabela 5. Uma combinação de 25% de *back-translation*, 25% de *part-of-speech* e 25% de *replace* e uma configuração com 1/3 de contribuição por cada método. Essas combinações não resultaram em melhorias significativas em relação aos resultados obtidos anteriormente. Mesmo com a combinação de diversas técnicas de aumento de dados, a quantidade de sentenças ainda é um fator limitante para alcançar melhorias de desempenho.

Tabela 5: Resultado em BLEU para combinações envolvendo *back-translation*, substituição alinhada e substituição por sinônimos

	Baseline	**⅓ Back-Translation + ⅓ Replace + ⅓ Part of Speech**	**½ Back-Translation + ¼ Replace + ¼ Part of Speech**
Básicas	42,05	54,24	52,80
Cardinais	50,18	63,44	63,78
Contexto	24,37	43,20	44,02
Direcionalidade	0,00	22,35	16,92
Famosos	27,34	49,94	50,71
Intensidade	38,47	42,44	33,28
Lugares	44,04	51,37	51,24
Negação	44,04	39,94	52,71
Romanos	25,86	60,28	62,68
Média	32,93	47,46	**47,57**

Figura 5: Média em BLEU do desempenho das técnicas de *data augmentation* com limitação de 155 mil sentenças no conjunto de treinamento.



6. Considerações finais

A geração de dados sintéticos para aprimorar modelos de Tradução Automática Neural em cenários de poucos recursos (*low-resource*), especialmente para línguas de sinais como a LIBRAS, é de suma importância para a acessibilidade e inclusão de pessoas surdas. Este trabalho explorou diferentes métodos de *data augmentation* a fim de identificar quais métodos trariam uma melhora no desempenho do tradutor segundo a métrica BLEU.

Observou-se que a técnica amplamente utilizada de *back-translation* também é eficaz na tradução de Português para LIBRAS, trazendo um ganho de 15,82 pontos de BLEU em relação ao *baseline*. A combinação de *back-translation* e substituição por sinônimos com *part-of-speech tagging* trouxe os melhores resultados em ambos os cenários: sem restrição no tamanho do conjunto de treinamento (+19,25 BLEU sobre o *baseline*) e também quando o conjunto de treinamento foi limitado a 155 mil sentenças (+16,5 BLEU sobre o *baseline*). Essas técnicas podem ser aplicadas para aumentar a quantidade de exemplos em conjuntos de sentenças que estejam sub-representados no corpus. Os resultados também confirmam a importância de manter um equilíbrio entre a quantidade de dados sintéticos gerados em relação à quantidade de dados autênticos no corpus original.

A investigação de técnicas mais custosas, porém potencialmente mais eficazes, pode abrir novas possibilidades para aprimorar ainda mais a qualidade da tradução automática. Modelos de linguagem têm demonstrado capacidade de produzir texto fluente em uma variedade de contextos linguísticos. Utilizar esses modelos para gerar dados sintéticos é uma sugestão para trabalhos futuros.

Referências

Veríssimo, V., Silva, C., Hanael, V., Moraes, C., Costa, R., Maritan, T., Aschoff, M., & Gaudêncio, T. A study on the use of sequence-to-sequence neural networks for automatic translation of Brazilian Portuguese to LIBRAS. In **Proceedings of the 25th Brazilian Symposium on Multimedia and the Web** (pp. 101–108). Rio de Janeiro, Brazil: Association for Computing Machinery, 2019.

ARAÚJO, T. M. U. **Uma solução para geração automática de trilhas em língua brasileira de sinais em conteúdos multimídia.** Tese (Doutorado em Automação e Sistemas) - Universidade Federal do Rio Grande do Norte, Natal, 2012. Disponível em: https://repositorio.ufrn.br/handle/123456789/15190. Acesso em: 8 abr. 2024

COSTA, Renan Paiva Oliveira *et al.* Avaliação do uso de modelos de aprendizagem profunda na tradução automática de línguas de sinais. **Revista Principia - Divulgação Científica e Tecnológica do IFPB**, João Pessoa, jan. 2024. ISSN 2447-9187. Disponível em: <https://periodicos.ifpb.edu.br/index.php/principia/article/view/8053/2488>. Acesso em: 09 Abr. 2024. doi:http://dx.doi.org/10.18265/2447-9187a2022id8053.

Poncelas *et al.* Investigating Backtranslation in Neural Machine Translation. In **Proceedings of the 21st Annual Conference of the European Association for Machine Translation** (pp. 269–278). Alicante, Spain, mai. 2018. https://aclanthology.org/2018.eamt-main.25

WANG, Xinyi *et al*. *SwitchOut: an Efficient Data Augmentation Algorithm for Neural Machine Translation.* In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, p.856-861, nov. 2018.

FENG, S. Y. *et al.* A Survey of Data Augmentation Approaches for NLP. In: **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021.** Online: Association for Computational Linguistics, 2021. p. 968–988.

FADAEE, Marzieh; BISAZZA, Ariana; MONZ, Christof. Data Augmentation for Low-Resource Neural Machine Translation. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)**, v. 2, p. 567–573, jul. 2017.

SÁNCHEZ-CARTAGENA, Víctor M. *et al*. Rethinking Data Augmentation for Low-Resource Neural Machine Translation: A Multi-Task Learning Approach. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, p.8502-8516, nov. 2021.

MAIMAITI, M. *et al.* Improving data augmentation for low-resource NMT guided by POS-tagging and paraphrase embedding. **Transactions on Asian and Low-Resource Language Information Processing**, v. 20, n. 6, p. 1-21, 2021.

JANG, Jin Yea *et al.* Automatic gloss-level data augmentation for sign language translation. In: **Proceedings of the Thirteenth Language Resources and Evaluation Conference.** 2022. p. 6808-6813.

WANG, Jing; YANG, Lina. Effective Data Augmentation Methods for CCMT 2022. In: **China Conference on Machine Translation.** Singapore: Springer Nature Singapore, 2022. p. 135-142.

VASWANI, Ashish *et al.* (2017). Attention is all you need. In: **Proceedings of the 31st International Conference on Neural Information Processing Systems.** Long Beach, California, USA: Curran Associates Inc. p. 6000-6010, 2017.

RANATHUNGA, S. *et al*. Neural Machine Translation for Low-resource Languages: A Survey. **ACM Comput. Surv**., New York, v. 55, n. 11, nov. 2023. ISSN 0360-0300.