XX ENCONTRO NACIONAL DE ENGENHARIA E DESENVOLVIMENTO SOCIAL
Construindo uma Engenharia Decolonial para a Soberania Digital e Popular
29 a 31 de outubro de 2025
Campinas - SP, Brasil

Tradução Automática da Língua Brasileira de Sinais

Leonardo Rener de Oliveira, Unicamp, l201270@dac.unicamp.br
José Mario de Martino, Unicamp, martino@unicamp.br

ARTIGO
EIXO TEMÁTICO: Engenharia, acessibilidade e tecnologias assistivas

RESUMO

A comunicação entre surdos usuários da Língua Brasileira de Sinais (Libras) e ouvintes não sinalizadores enfrenta barreiras significativas e a tradução automática de língua de sinais (SLT) apresenta desafios como a variabilidade dos gestos e a modelagem espaço-temporal. Este artigo propõe um sistema de tradução de sentenças contínuas em Libras para o português escrito, utilizando uma arquitetura Transformer em abordagem end-to-end. Utilizou-se o conjunto de dados PorLibras, com vídeos convertidos em sequências de poses 3D, sem glosas intermediárias. Os resultados foram expressivos, com BLEU-4 de 77,33% e WER de 30,74%. A principal contribuição é demonstrar a viabilidade de um modelo Transformer robusto para SLT, avançando na tradução automática de Libras e promovendo maior inclusão da comunidade surda.

PALAVRAS-CHAVE: Tecnologia Assistiva. Língua Brasileira de Sinais (Libras). Tradução Automática. Reconhecimento de Língua de Sinais. Processamento de Linguagem Natural (PLN).

INTRODUÇÃO

A comunicação é um pilar fundamental da interação social e do acesso à informação. No Brasil, cerca de 10,7 milhões de pessoas com mais de cinco anos possuem algum grau de deficiência auditiva, e uma parcela de 22,4% delas utiliza a Língua Brasileira de Sinais (Libras) como principal meio de comunicação. Um dos maiores desafios enfrentados por essa comunidade é a escassez de ouvintes fluentes em Libras, o que pode transformar atividades cotidianas em barreiras significativas, restringindo a participação plena na sociedade. Nesse cenário, as tecnologias assistivas (TA) surgem como ferramentas com grande potencial para mitigar essas barreiras, promovendo a autonomia e melhorando a qualidade de vida de quem se comunica por Libras.

A Tradução Automática de Língua de Sinais (SLT, do inglês Sign Language Translation) é um campo de pesquisa em constante evolução, focado na conversão de vídeos de sinais em texto ou fala. A tarefa apresenta desafios únicos devido à natureza visual-gestual das línguas de sinais, que utilizam componentes manuais, como configuração e movimento das mãos, e não manuais, como expressões faciais e movimentos do tronco, para construir o significado. Os principais desafios técnicos envolvem a extração robusta de características visuais, a modelagem das complexas dinâmicas espaço-temporais dos sinais e a tradução de sequências contínuas, que apresentam alta variabilidade individual e contextual.

A literatura recente destaca duas abordagens principais para a SLT: a que divide o processo em duas etapas (reconhecimento de sinais para glosas e depois tradução de glosas para texto) e a abordagem end-to-end, que treina um único modelo para converter o vídeo diretamente em texto. Nos últimos anos, a arquitetura Transformer, introduzida por Vaswani et al., tornou-se o padrão em muitas tarefas de processamento de linguagem natural por sua eficiência em capturar dependências de longo prazo

através de mecanismos de autoatenção. Apesar do avanço em línguas como a Americana (ASL), há uma carência de estudos e de conjuntos de dados para a tradução de sentenças contínuas em Libras, com muitos trabalhos focados apenas no reconhecimento de sinais isolados.

Este trabalho aborda essa lacuna, propondo o desenvolvimento e a avaliação de um modelo de tradução automática, baseado na arquitetura Transformer, para a conversão de sentenças contínuas da Língua Brasileira de Sinais (Libras) para o português escrito. Embora o objetivo final seja integrar este modelo a um sistema de Realidade Aumentada (RA) para a exibição de legendas em óculos inteligentes, o escopo deste artigo concentra-se exclusivamente no desenvolvimento e na validação do módulo de tradução. Portanto, os objetivos específicos deste artigo são: Desenvolver um modelo de tradução automática de Libras para o português com base na arquitetura Transformer, operando diretamente sobre dados de pose extraídos dos vídeos; Avaliar a qualidade e a precisão das traduções geradas pelo modelo por meio de métricas padronizadas na área, como BLEU, METEOR e WER; Contribuir com a pesquisa em tecnologia assistiva para a comunidade surda brasileira, fornecendo uma análise do desempenho de um modelo de ponta para a tradução de Libras.

METODOLOGIA

Esta pesquisa propõe o desenvolvimento e a avaliação de um sistema de tradução automática para a Libras, que integra uma arquitetura Transformer para realizar a tradução de maneira end-to-end. A metodologia abrange a seleção da base de dados, o pré-processamento para representação dos sinais, a arquitetura do modelo de tradução e o processo de treinamento, conforme detalhado a seguir.

Base de dados

A qualidade e a quantidade dos dados são fatores determinantes para o desenvolvimento de sistemas de aprendizado de máquina robustos. Para este trabalho, foi realizada uma busca por bases de dados de Libras que contivessem vídeos de sentenças contínuas acompanhadas de suas respectivas traduções para o português. Dentre as opções, foi selecionado o corpus PorLibras , que é composto pela tradução de um livro escolar. Este conjunto de dados contém mais de 2.000 frases, totalizando aproximadamente 8 horas de material audiovisual, e inclui anotações precisas de glosas, o que o torna um recurso valioso para a pesquisa em SLT.

Pré-processamento e representação de dados

O processo de tradução inicia-se com a extração de características das gravações de vídeo. Para isso, foi utilizado o framework MediaPipe para extrair pontos-chave de pose 3D de cada quadro. Inicialmente, são capturados 543 pontos por quadro, englobando rosto, pose e ambas as mãos.

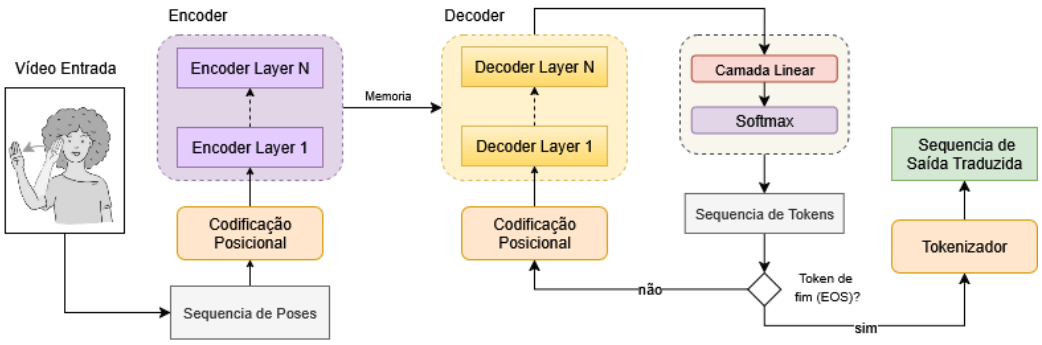
Visando reduzir a complexidade e focar nos dados mais relevantes para a língua de sinais, o número de pontos-chave foi reduzido para 178 por quadro. Foram preservados os pontos relacionados às mãos, ombros, cotovelos, pulsos, quadris e os principais pontos do rosto (contorno facial, lábios, olhos e nariz). Isso resultou em um vetor de características com 534 dimensões para cada quadro do vídeo. Para a manipulação e aumento dos dados de pose, foi utilizada a biblioteca pose-format, aplicando técnicas como rotação, espelhamento e adição de ruído para melhorar a capacidade de generalização do modelo.

Arquitetura do modelo

A arquitetura segue o paradigma clássico de codificador-decodificador, composto por três blocos principais: um codificador Transformer especializado para processar sequências de poses (Encoder), um decodificador Transformer para geração de texto

(Decoder), e uma camada de projeção linear que converte os vetores do decodificador em distribuições de probabilidade sobre o vocabulário. A Figura 1 apresenta o fluxo completo do modelo implementado.

Figura 1 - Arquitetura do modelo, com abordagem end-to-end.



Fonte: Elaboração própria.

O codificador recebe como entrada uma sequência de vetores representando poses corporais. Cada vetor de entrada é primeiro transformado por uma camada linear para projetar os dados para um espaço vetorial de dimensão compatível com o modelo. Em seguida, é aplicado um codificador posicional (PositionalEncoding) que insere informações temporais na sequência. O bloco codificador é composto por múltiplas camadas do tipo TransformerEncoderLayer, com mecanismos de autoatenção multi-cabeça (multi-head self-attention) e redes feedforward posicionais.

O decodificador recebe como entrada a memória do Encoder, juntamente com a sequência alvo (durante o treinamento) ou a sequência previamente gerada (durante a inferência). Inicialmente, os tokens são convertidos em embeddings por uma camada de Embedding, seguidos por codificação posicional e múltiplas camadas TransformerDecoderLayer, que combinam autoatenção na sequência alvo com atenção cruzada à memória gerada pelo codificador.

Ao término do processo de decodificação, uma camada linear projeta os vetores de saída em uma distribuição de probabilidade sobre o vocabulário. Essa projeção é seguida por uma função softmax, que atribui uma probabilidade a cada palavra candidata, permitindo a geração sequencial (autoregressiva) dos tokens que compõem a legenda

final. O token com a maior probabilidade é selecionado e anexado à sequência de tokens. Este processo se repete até que o token <EOS> seja gerado. A sequência final de tokens é então convertida em texto pelo Tokenizador.

Treinamento do modelo

O treinamento foi conduzido em um sistema com processador Intel Core i7-9700KF e uma placa de vídeo RTX 2080 Ti. O modelo foi treinado por 30 épocas, totalizando aproximadamente 9 horas de execução. Para os experimentos, a arquitetura foi configurada com 3 camadas, 8 cabeças de atenção e uma dimensão de 512 para as incorporações (embeddings).

Durante o treinamento, foi empregada a técnica de teacher forcing, na qual a sequência de texto verdadeira é fornecida como entrada para o decodificador, o que acelera a convergência e estabiliza o aprendizado. A otimização conjunta do codificador e do decodificador é realizada através do cálculo da perda de entropia cruzada, permitindo que o modelo aprenda a mapear o espaço contínuo das poses para o espaço discreto do texto de forma coerente.

Métricas de Avaliação

Para avaliar o desempenho do modelo de tradução de forma quantitativa e objetiva, foram selecionadas três métricas consagradas na área de tradução automática: BLEU, METEOR e WER. A combinação dessas métricas permite uma análise multifacetada, avaliando desde a precisão lexical e gramatical até a semelhança estrutural com a tradução de referência.

BLEU (Bilingual Evaluation Understudy): É uma métrica baseada em precisão que calcula a sobreposição de n-gramas (sequências de palavras) entre a tradução gerada pelo modelo e a tradução de referência. Neste trabalho, foram utilizadas as variações de BLEU-1 a BLEU-4 para medir a adequação da tradução em diferentes níveis, desde a

correspondência de palavras individuais (BLEU-1) até a fluidez de frases mais longas (BLEU-4).

METEOR (Metric for Evaluation of Translation with Explicit ORdering): Diferente do BLEU, o METEOR avalia a tradução com base em um alinhamento explícito entre as palavras da sentença gerada e da referência, considerando não apenas correspondências exatas, mas também sinônimos e radicais. A métrica calcula a precisão e o recall, aplicando uma penalidade por fragmentação para garantir que a ordem das palavras seja coerente, resultando em uma avaliação mais alinhada à percepção humana de qualidade.

WER (Word Error Rate): Com origem na área de reconhecimento de fala, a Taxa de Erro de Palavra é uma métrica que calcula a distância entre a sentença gerada e a referência. Ela mede o número mínimo de edições (substituições, deleções e inserções) necessárias para que a tradução do modelo se torne idêntica à referência. Um valor de WER mais baixo indica uma tradução de maior qualidade, com menos erros.

DESENVOLVIMENTO (RESULTADOS E DISCUSSÕES)

Após a conclusão da etapa de treinamento do modelo, foi realizada uma avaliação quantitativa para mensurar a qualidade das traduções geradas. Os resultados obtidos no conjunto de testes estão apresentados na Tabela 1.

Tabela 1 - Desempenho do modelo de tradução

Métrica	Valor Médio
BLEU-1	78,07
BLEU-2	77,68
BLEU-3	77,48
BLEU-4	77,33
METEOR	67,80
WER	30,74

Fonte: Elaboração própria.

A análise dos resultados revela um desempenho promissor do modelo. Os escores de BLEU, que medem a sobreposição de n-gramas entre a tradução gerada e a

referência, apresentam uma queda gradual à medida que se consideram n-gramas maiores (de BLEU-1 a BLEU-4). Esse comportamento é típico em tarefas de tradução de sinais e sugere que, embora o modelo seja eficaz na tradução de palavras e termos curtos (unigramas e bigramas), ele enfrenta desafios na captura de dependências gestuais de maior contexto, que se refletem em estruturas sintáticas mais longas e complexas.

O valor de METEOR, de 67,80%, corrobora a qualidade da tradução, pois essa métrica considera não apenas a precisão, mas também a ordem e a sinonímia das palavras, fornecendo uma avaliação mais alinhada à percepção humana. A Taxa de Erro de Palavra (WER) de 30,74% indica que, em média, cerca de 30% das palavras precisam ser ajustadas (substituídas, inseridas ou deletadas) para que a tradução gerada corresponda à referência. Embora haja espaço para melhorias, este é um resultado significativo para uma abordagem end-to-end que lida com a alta variabilidade da sinalização contínua em Libras.

CONSIDERAÇÕES FINAIS

Os resultados demonstram a viabilidade e a eficácia da abordagem. O modelo alcançou um desempenho quantitativo robusto, com um escore BLEU-4 de 77,33% e METEOR de 67,80%, indicando sua capacidade de gerar traduções coerentes e precisas para um problema de alta complexidade. A principal contribuição deste estudo é a aplicação e validação de um modelo de ponta para a tradução contínua de Libras, um campo com carência de pesquisas e recursos em comparação com outras línguas de sinais. Este trabalho representa, portanto, um passo importante para o avanço de tecnologias assistivas, com potencial real de melhorar a qualidade de vida e a independência de indivíduos surdos que se comunicam por Libras.

Apesar dos resultados promissores, é fundamental reconhecer as limitações do estudo para direcionar futuras pesquisas. Uma limitação central reside no viés do conjunto de dados utilizado. O corpus PorLibras é composto pela tradução de um livro

escolar, o que restringe o vocabulário e as estruturas frasais a um domínio específico. Consequentemente, o modelo pode não generalizar de forma eficaz para outros contextos, como conversas informais do dia a dia, que apresentam maior variabilidade e nuances regionais e culturais. Além dessa questão, como observado na análise dos escores de BLEU, a captura de dependências gestuais de maior contexto permanece um desafio técnico a ser aprimorado.

Para trabalhos futuros, além dos aperfeiçoamentos nos hiperparâmetros do modelo, destaca-se a necessidade da criação ou utilização de conjuntos de dados que abranjam múltiplos domínios e um maior número de sinalizadores, visando maior robustez. A exploração de técnicas como o aprendizado por transferência (transfer learning) também se apresenta como um caminho promissor para mitigar o viés dos dados e aprimorar a capacidade de generalização do sistema.

REFERÊNCIAS

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). Pesquisa Nacional de Saúde (PNS). 2019. Disponível em: <https://sidra.ibge.gov.br/pesquisa/pns/pns-2019>. Acesso em: 25/02/2025.

Westin, R. Baixo alcance da língua de sinais leva surdos ao isolamento. 2019. **Senado Notícias**. Disponível em: <https://www12.senado.leg.br/noticias/especiais/especial-cidadania/baixo-alcance-da-lingua-de-sinais-leva-surdos-ao-isolamento>. Acesso em: 25/02/2025.

WANG, H. et al. Progress in Machine Translation. **Engineering**, v. 18, 14 jul. 2021.

SAUNDERS, B.; NECATI CIHAN CAMGÖZ; BOWDEN, R. Progressive Transformers for End-to-End Sign Language Production. **Lecture Notes in Computer Science**, p. 687–705, 1 jan. 2020.

VASWANI, A. et al. Attention is All You Need. **Advances in Neural Information Processing Systems**, v. 30, p. 5998–6008, 2017.

CAMGOZ, N. C. et al. Neural Sign Language Translation. **2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition**, jun. 2018.

Johnston, T. et al. Corpus linguistics and signed languages: no lemmata, no corpus. **3RD Workshop on the Representation and Processing of Sign Languages**. 2008.

NECATI CIHAN CAMGÖZ et al. Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. **arXiv (Cornell University)**, 30 mar. 2020.

KO, S.-K. et al. Neural Sign Language Translation Based on Human Keypoint Estimation. **Applied Sciences**, v. 9, n. 13, p. 2683, 1 jan. 2019.

LUGARESI, C. et al. MediaPipe: A Framework for Building Perception Pipelines. **arXiv**, 14 jun. 2019.

De Martino, J. M. et al. Building a Brazilian Portuguese-Brazilian sign language parallel corpus using motion capture data. **THE 12th International Conference on the Computational Processing of the Portuguese Language, Tomar. Proceedings Workshop on Corpora and Tools for Processing Corpora Workshop**, p. 56–63, jul. 2016.

MORYOSSEF, A.; MÜLLER, M.; FAHRNI, R. pose-format: Library for Viewing, Augmenting, and Handling .pose Files. **arXiv (Cornell University)**, 1 jan. 2023.

PAN, S. J.; YANG, Q. A Survey on Transfer Learning. **IEEE Transactions on Knowledge and Data Engineering**, v. 22, n. 10, p. 1345–1359, out. 2010.