Restauração de Pontuação em Textos Traduzidos no Idioma pt-BR a partir de Transcrição de Áudios

Angel G. de S. Sales[1], **Brenda C. D. Moura**[1],
José E. B. de S. Linhares[1], **Fabiann M. D. Barbosa**[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Amazonas (IFAM)
Campus Manaus Zona Leste – Manaus, Amazonas – Brasil

{angelgabrieldesouzasales, brendacdmoura}@gmail.com,

{breno.linhares,fabiann.dantas}@ifam.edu.br

Abstract. *This work proposes a model for automatically restoring punctuation in colloquial Brazilian Portuguese texts derived from audio transcriptions, focusing on improving readability and usability in NLP tasks. The methodology involves two main stages: training and inference. The IWSLT (2014–2016) corpus is used, which contains translated TED talk transcriptions, and a hybrid model composed of Bi-LSTM, self-attention, and CRF. Preprocessing includes punctuation mapping, vocabulary construction, GloVe embeddings, and highway network. Four evaluation scenarios were applied, revealing that combining the three years of data yields the best results, with progressive improvements in precision, recall, and F1-score metrics.*

Resumo. *Este trabalho propõe um modelo para restaurar automaticamente a pontuação em textos coloquiais em português do Brasil, originados de transcrições de áudio, com foco na melhoria da legibilidade e usabilidade em tarefas de PLN. A metodologia envolve duas etapas principais: treinamento e inferência. Utiliza-se o corpus IWSLT (2014–2016), que contém transcrições de palestras TED traduzidas, e um modelo híbrido com Bi-LSTM, self-attention e CRF. O pré-processamento inclui mapeamento de sinais, construção de vocabulário, embeddings GloVe e rede highway. Quatro cenários de avaliação foram aplicados, revelando que a combinação dos dados dos três anos gera os melhores resultados, com melhorias progressivas nas métricas de precisão, recall e F1-score.*

1. Introdução

Por meio da linguagem, os seres humanos expressam ideias, emoções e intenções de forma complexa. Embora sistemas computacionais ainda não consigam emular integralmente essa habilidade, avanços no Processamento de Linguagem Natural (PLN) têm permitido que máquinas compreendam e gerem linguagem natural a partir de grandes volumes de dados e conjuntos de referência estáveis [Olive et al. 2011, Caseli and Nunes 2023].

O PLN busca reduzir a distância entre a linguagem humana e a computacional, enfrentando desafios como variações sintáticas, dialetos e contextos culturais. Um exemplo relevante é o Reconhecimento Automático de Fala (RAF), cujas transcrições

geralmente não contêm pontuação, o que pode comprometer a legibilidade e afetar tarefas *downstream*, como tradução automática e análise semântica [Guerreiro et al. 2021, Chordia 2021].

Nesse contexto, a restauração automática de pontuação é uma etapa importante no pós-processamento de textos gerados por sistemas automáticos, com o objetivo de inserir sinais ausentes com base no contexto linguístico [Lima et al. 2022]. No entanto, há uma concentração de estudos em idiomas como o inglês, com menor cobertura voltada ao português do Brasil, especialmente em contextos informais [de Lima et al. 2024, de Lima et al. 2023, Gris et al. 2023].

Este trabalho propõe uma abordagem para restaurar a pontuação em textos coloquiais em pt-BR, a partir de transcrições de áudio, utilizando uma arquitetura neural híbrida com *embeddings GloVe*, rede *highway*, Bi-LSTM, *self-attention* e CRF, treinada com dados do IWSLT (2014–2016), que contém transcrições de palestras TED traduzidas. A proposta busca contribuir com a pesquisa em PLN e apoiar aplicações práticas, como corretores gramaticais, legendas automáticas, tradução e agentes conversacionais.

O artigo está estruturado da seguinte maneira: a Seção 2 revisa os principais trabalhos relacionados; a Seção 3 descreve a metodologia proposta; a Seção 4 apresenta o *setup* e os resultados; por fim, a Seção 5 discute os achados e propõe direções para trabalhos futuros.

2. Trabalhos Relacionados

A restauração de pontuação em transcrições de fala é uma tarefa essencial no PLN, com impacto direto na inteligibilidade do texto e na performance de tarefas, como tradução automática e análise sintática. Nos últimos anos, esse problema tem sido enfrentado com o uso de técnicas baseadas em aprendizado profundo e mais recentemente, modelos de linguagem de grande escala (LLMs).

Inicialmente, métodos como o proposto por [Chordia 2021] adotaram abordagens com *embeddings* de palavras e redes recorrentes bidirecionais, associadas a decodificadores CRF, obtendo resultados relevantes em idiomas de alto e baixo recurso. Esse tipo de arquitetura ainda é recorrente na literatura por sua capacidade de modelar sequências.

No cenário do idioma português, poucos estudos se destacam. O trabalho de [Lima et al. 2022] foi um dos primeiros a explorar comparativamente diferentes modelos para a restauração de pontuação em pt-BR. Eles avaliaram Bi-LSTM com CRF e BERT, obtendo desempenho superior com o segundo, embora com queda significativa ao testar domínios distintos (*cross-domain*). Ademais, também é aplicado um estudo de performance evidenciando a necessidade de alta capacidade de processamento para utilização da arquitetura BERT.

Avançando em direção ao uso mais explícito de LLMs, [de Lima et al. 2023] avaliou os modelos pré-treinados BERT e T5 em textos educacionais em português, observando ganhos importantes de desempenho, especialmente em domínios específicos, onde a arquitetura T5 alcançou 89% de *f1-score* geral. O estudo evidencia o potencial de generalização e especialização dessas arquiteturas.

Mais recentemente, [Gris et al. 2023] investigaram a aplicação do *Whisper*, modelo da OpenAI para reconhecimento de fala, na tarefa de restauração de pontuação em

português do Brasil. Os autores identificaram desafios notáveis na acurácia de pontuações menos frequentes e destacaram a necessidade de ajustar esses sistemas às características da língua portuguesa.

Em contrapartida, [Moura et al. 2025] destaca a importância de investigar arquiteturas com menor custo computacional, uma vez que, conforme exposto por [Lima et al. 2022], modelos como o BERT demandam mais recursos, dificultando sua aplicação em contextos de menor escala. Nesse sentido, o estudo propõe a utilização da arquitetura Bi-LSTM + *self-attention* + CRF, aplicada em um cenário de linguagem formal e com validação *cross-domain* por meio do *dataset* IWSLT. No entanto, observou-se que, devido às diferenças inerentes entre sentenças formais e informais, os resultados obtidos na análise *cross-domain* foram insatisfatórios.

Apesar desses avanços, ainda há lacunas importantes: a maioria dos trabalhos utiliza corpora limitados, não trata adequadamente a linguagem coloquial e informal (frequente em contextos reais), e pouco explora o potencial combinatório de diferentes mecanismos de atenção, *embeddings* e estruturas sequenciais para a criação de modelos balanceados em questão de performance e precisão.

A proposta deste trabalho visa preencher essas lacunas ao empregar uma arquitetura híbrida composta por Bi-LSTM, *self-attention* e CRF, aplicada especificamente a transcrições traduzidas para o português do Brasil, com foco em linguagem coloquial. Uma visão consolidada das principais diferenças entre os trabalhos analisados é apresentada na Tabela 1, a qual compara as abordagens utilizadas, os idiomas, os corpora empregados e as contribuições específicas de cada estudo.

Tabela 1. Comparação entre os trabalhos relacionados

Trabalho	**Ano**	**Modelo**	**Idioma**	**Corpus**	**Pontuação Avaliada**	**Contribuição Principal**
Chordia (2021)	2021	Bi-LSTM + CRF	Multilíngue	Discursos políticos, artigos de notícias	Vírgula, ponto, interrogação	Arquitetura robusta baseada em RNN e CRF para diversos idiomas
Lima et al. (2022)	2022	Bi-LSTM + CRF vs. BERT	pt-BR	IWSLT 2012, OBRAS (*cross-domain*)	Vírgula, ponto, interrogação	Avaliação comparativa entre modelos clássicos e BERT para pt-BR
Lima et al. (2023)	2023	BERT, T5	pt-BR	Textos educacionais	Diversos sinais	Avaliação de LLMs no contexto educacional brasileiro
Gris et al. (2023)	2023	*Whisper* (OpenAI)	pt-BR	Transcrições de fala	Vírgula, ponto, interrogação	*Benchmark* de modelo automático de fala para pontuação
Moura et al. (2025)	2025	Bi-LSTM + CRF + *Self-attention*	Português	*Portuguese Legal Sentences v3* e IWSLT	Vírgula, ponto, interrogação	Treinamento com textos formais e avaliação *cross-domain* em dados informais.
Este trabalho	2025	Bi-LSTM + CRF + *Self-Attention*	pt-BR	IWSLT (2014–2016), textos traduzidos	Vírgula, ponto, interrogação	Arquitetura híbrida com análise de evolução temporal e linguagem coloquial

3. Metodologia

Nesta seção, é apresentada a abordagem utilizada para restauração de pontuação em textos traduzidos para o idioma português brasileiro (pt-BR) a partir da transcrição de áudios originalmente em inglês. A metodologia empregada foi fundamentada no pro-

jeto *Punctuation-Restoration*, desenvolvido por Luo[1]. A seleção deste projeto como base foi decorrente de sua arquitetura robusta e comprovadamente eficaz para tarefas de restauração de pontuação em sequências textuais. Adicionalmente, a disponibilidade pública do código-fonte do projeto de Luo facilitou a replicabilidade e a adaptação das técnicas para o contexto específico do pt-BR, permitindo que a pesquisa se concentrasse na avaliação e otimização do modelo para o idioma alvo. Consequentemente, a escolha das métricas de avaliação da arquitetura do modelo e das técnicas aplicadas nesta pesquisa derivou diretamente deste projeto, que serviu como base referencial sólida

O sistema proposto é composto por duas etapas principais: (i) o treinamento do modelo de restauração de pontuação, e (ii) a etapa de inferência, na qual o modelo treinado é utilizado para predizer a pontuação em novas transcrições. A seguir, apresenta-se o detalhamento do diagrama em blocos da metodologia proposta, que representa o funcionamento geral do sistema, conforme ilustrado na Figura 1.

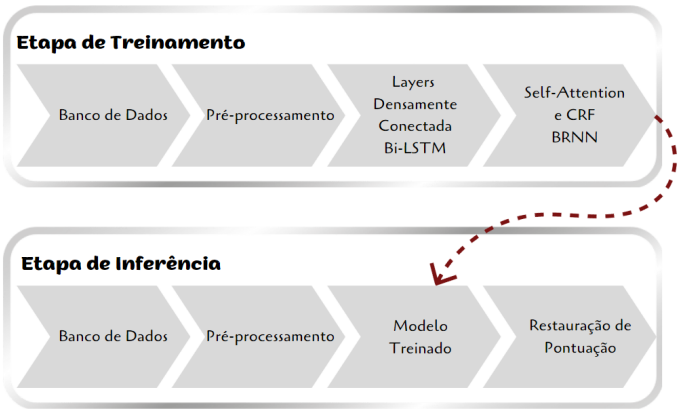

Figura 1. Diagrama em blocos da metodologia proposta.

A presente pesquisa adota uma metodologia de caráter exploratório e descritivo. O aspecto exploratório se justifica pela limitada quantidade de investigações direcionadas à restauração de pontuação em textos de transcrição automática em pt-BR, indicando lacunas ainda presentes na literatura. Já o caráter descritivo visa demonstrar como os diversos sinais de pontuação se manifestam no corpus e explicitar a maneira pela qual o modelo proposto processa e se adapta a essas particularidades linguísticas, delineando seu comportamento em termos de desempenho e aprendizado ao longo dos cenários de avaliação.

Adicionalmente, a pesquisa emprega uma abordagem mista, combinando elementos quantitativos e qualitativos. A vertente quantitativa possibilita a mensuração objetiva e o rigoroso acompanhamento do desempenho dos modelos por meio de métricas estatísticas. A dimensão qualitativa, por sua vez, manifesta-se explicitamente na análise interpretativa dos insights derivados do modelo, na compreensão detalhada de seu comportamento diante das particularidades da linguagem (por exemplo, os desafios enfrentados com pontuações de baixa frequência), e na discussão das implicações dos resultados obtidos, o que permite delinear lacunas na pesquisa e propor futuras direções de estudo.

[1]Link do projeto de referência disponível em: `https://github.com/k9luo/Punctuation-Restoration`

3.1. Base de dados

Na etapa de base de dados, o estudo empregou o *dataset* IWSLT (*International Workshop on Spoken Language Translation*), abrangendo os anos de 2014, 2015, e 2016. Este conjunto de dados, proveniente do WIT3 (*Web Inventory of Transcribed and Translated Talks*) , consiste em transcrições de palestras TED, originalmente em inglês e traduzidas para o português. A escolha do IWSLT é estratégica, pois, ao conter linguagem próxima da fala espontânea e coloquial, representa um corpus linguisticamente adequado para problemas de Restauração Automática de Pontuação (RAP) em um domínio informal. Apesar de sua natureza bilíngue, a pesquisa focou exclusivamente no conteúdo em português, descartando os segmentos em inglês, visando fomentar o avanço do PLN para o português brasileiro, um campo que ainda demanda soluções eficazes. É importante ressaltar que este conjunto de dados não constitui uma base acumulativa, pois os dados variam entre os anos mencionados.

A seleção de transcrições de anos distintos teve como objetivo analisar como a evolução temporal dos dados impacta o desempenho do modelo. O uso combinado de diferentes anos busca não apenas aumentar o volume de dados, o que tende a melhorar a capacidade de generalização e reduzir o *overfitting*, mas também ampliar a diversidade linguística. Com isso, considera-se a possibilidade de que mudanças nos temas das palestras e na linguagem coloquial ao longo do tempo demandem modelos mais adaptáveis e sofisticados. Além disso, a seleção foi condicionada pela disponibilidade dos dados, sempre mantendo o compromisso com a construção de uma base linguística mais madura e representativa da oralidade em português brasileiro.

Para avaliar o desempenho do modelo treinado utilizando a base de dados IWSLT, foram estabelecidos quatro cenários: no Cenário 1, utilizou-se o conjunto referente ao ano de 2014; no Cenário 2, utilizou-se o conjunto referente ao ano de 2015; no Cenário 3, utilizou-se o conjunto referente ao ano de 2016; e no Cenário 4, utilizou-se a combinação dos conjuntos referentes aos anos de 2014 a 2016. Observa-se na Tabela 2 que, para cada ano em que o conjunto foi produzido, existe um aumento no volume dos textos em termos do número de palavras. Além disso, verifica-se uma predominância de pontuação a nível de vírgula em relação ao ponto final e ponto de interrogação. Para cada conjunto referente aos anos de 2014 a 2016, existe um subconjunto para treino (*Train* e Dev) e teste (Ref e ASR) do modelo.

Tabela 2. Comparação das estatísticas do conjunto *Train* nos cenários 1 a 4.

Cenário	*Dataset*	*Total Number of Words*	*Number of Unique Words*	*Number of Comma*	*Number of Period*	*Number of Question Mark*
1	Train	67655	14515	3275	2578	352
2	Train	87706	17235	4483	565	3430
3	Train	100781	18654	3955	5293	652
4	Train	256682	21762	12971	9971	1520

3.2. Pré-processamento

Na etapa de pré-processamento, foram aplicadas estratégias para preparar a base de dados para o treinamento do modelo. Inicialmente, foi implementado um mapeamento para o tratamento de diferentes sinais de pontuação, no qual a vírgula, o ponto e o ponto de interrogação foram convertidos em categorias do vocabulário de saída, enquanto os demais sinais foram removidos do corpus. O estudo concentrou-se exclusivamente nesses

três sinais, dada a sua relevância na construção de textos mais claros e compreensíveis. Também foi realizada a construção de um vocabulário contendo todas as palavras do conjunto de treinamento. Por fim, os textos foram segmentados em unidades menores, como frases ou trechos, com o objetivo de otimizar o processamento nas etapas seguintes.

Em seguida, utilizou-se o modelo de *word embeddings* GloVe (*Global Vectors for Word Representation*) com 300 dimensões, a fim de obter representações vetoriais das palavras. Esse modelo é capaz de capturar relações semânticas no espaço vetorial por meio de *embeddings* com pesos e dimensões específicas [Pennington et al. 2014]. Por fim, a arquitetura inclui uma rede *highway* com duas camadas, cuja função é facilitar o fluxo de informações e a aprendizagem em redes profundas por meio de mecanismos de *gating*, que controlam quais conexões são mais relevantes durante o processo de treinamento [Srivastava et al. 2015].

3.3. *Layers* Densamente Conectada Bi-LSTM

Nesta etapa, os *embeddings* são incorporados a uma rede neural com o objetivo de realizar a predição das pontuações. Para a construção dessa rede, foram utilizadas camadas densamente conectadas do tipo *Bidirectional Long Short-Term Memory* (Bi-LSTM), compondo a estrutura de uma rede neural recorrente. Especificamente, foram empregadas quatro camadas com 50, 50, 50 e 300 unidades, respectivamente.

3.4. *Self-Attention* e CRF BRNN

Na etapa *Self-Attention* e CRF BRNN, o mecanismo de *self-attention* foi implementado para permitir que o modelo direcione sua atenção para partes específicas do texto durante a previsão. Essa atenção, empregada tanto no treinamento quanto na inferência, capacita o modelo a focar em distintas seções do texto, aprimorando, assim, sua habilidade em identificar informações pertinentes.

O decodificador CRF é integrado à arquitetura para otimizar a sequencialização das pontuações restauradas, sendo utilizado na construção de camadas de RNN densamente conectadas. Ao longo da fase de treinamento, o decodificador CRF facilita a aprendizagem das dependências sequenciais entre as pontuações, contribuindo significativamente para o aprimoramento da precisão na restauração.

3.5. Restauração da Pontuação

Nessa fase, o modelo realiza previsões para restaurar a pontuação no texto, oferecendo uma visão clara de como os *tokens* previstos são representados. Os sinais de pontuação são codificados com base nas previsões do modelo, fornecendo uma compreensão detalhada de como as restaurações são organizadas e de como o modelo interpreta e aplica as pontuações.

Esta abordagem não apenas destaca a estrutura da saída do modelo, mas também apresenta métricas fundamentais, tais como: ERR (Erro Relativo de Restauração) e SER (Taxa de Erro de *Slot*). Essas métricas desempenham um importante papel na avaliação do desempenho na restauração de pontuação, apresentando *insights* valiosos sobre a eficácia do modelo, além de enriquecer a compreensão de seu comportamento durante a inferência.

Além de ERR e SER, utilizam-se também outras métricas de avaliação de desempenho, como precisão, *recall* e *f1-score*, com o intuito de se obter uma visão mais abrangente da eficácia do modelo na restauração de pontuação. Ademais, durante o treinamento de modelos de aprendizado de máquina, duas métricas se destacam na avaliação do desempenho: a *train loss* e a *perplexity*.

A primeira mede a disparidade entre a saída gerada pelo modelo e a saída esperada, com base nos dados de treinamento, sendo utilizada para ajustar os parâmetros e monitorar o progresso do treinamento. Já a *perplexity* é uma métrica voltada à avaliação da capacidade do modelo de linguagem em antecipar sequências textuais, sendo geralmente aplicada sobre dados de validação ou teste. Essa medida reflete, de forma direta, o grau de confiança do modelo em suas previsões.

4. Resultados e Discussão

Nesta seção, serão elencados os resultados obtidos no decorrer da pesquisa. Para uma melhor explanação, dividiu-se em sete partes: *Setup*, Treinamento e Teste dos Cenários 1 a 4 e finalizando, faz-se uma análise geral destes resultados.

4.1. *Setup*

Os experimentos realizados nesta pesquisa utilizaram a linguagem Python (versão `3.10.12`), com um *cluster* conforme especificado: SO Linux, na distribuição Ubuntu, versão `20.04.4` LTS; GPU NVIDIA Geforce RTX 3090, com 24 GB; e 24x CPU Intel(R) Core(TM) i9-12900K, 12th Gen.

4.2. Treinamento e Teste do Cenário 1

O gráfico de *loss* e *perplexity* (ver Figura 2) ilustra a evolução do modelo durante o treinamento. Observa-se que a *loss* diminui mais lentamente, sugerindo uma possível saturação do modelo em termos de generalização. Entretanto, a *perplexity* decai rápido, indicando aprendizado efetivo.

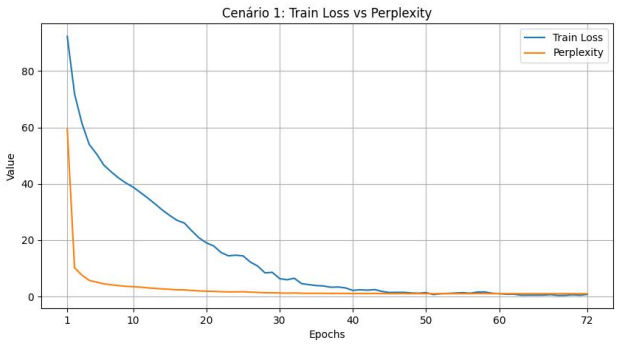

Figura 2. Gráfico de *Loss* e *Perplexity* do Cenário 1.

Os resultados apresentados na Tabela 3 indicam que a predição de vírgulas (`<comma>`) é mais eficaz em comparação com a predição de pontos (`<period>`). Os valores de ERR e SER refletem a capacidade do modelo em inserir pontuações de forma consistente e precisa nas transcrições.

Tabela 3. Resultados da avaliação nos conjuntos de dados `ref.txt` e `asr.txt`.

Dataset	Punctuation	Precision	Recall	F1-score
ref.txt	<comma>	97.63	97.86	97.75
	<period>	96.18	98.37	97.26
	<questionmark>	97.83	91.84	94.74
	Overall	97.06	97.68	97.37
	ERR	0.41%		
	SER	4.6%		
asr.txt	<comma>	99.40	98.14	98.77
	<period>	97.16	99.25	98.19
	<questionmark>	97.97	96.03	96.99
	Overall	98.36	98.49	98.43
	ERR	0.26%		
	SER	2.9%		

4.3. Treinamento e Teste do Cenário 2

Os resultados apresentados na Tabela 4 indicam melhorias em relação ao Cenário 1. A precisão, *recall* e *f1-score* geral demonstram avanços, especialmente na predição de vírgulas (<comma>). Entretanto, desafios persistem na predição de pontos de interrogação, especialmente em relação ao decréscimo de precisão. Os valores de ERR e SER também apresentam diminuição em comparação aos resultados do cenário anterior, o que é um indicador positivo de que o modelo está cometendo menos erros. Essa melhoria sugere que o modelo está se tornando mais eficiente na identificação correta das posições para inserção de pontuação, embora a complexidade da predição de pontos de interrogação ainda demande atenção específica.

Tabela 4. Resultados da avaliação nos conjuntos de dados `ref.txt` e `asr.txt`.

Dataset	Punctuation	Precision	Recall	F1-score
ref.txt	<comma>	97.69	99.22	98.45
	<period>	97.09	97.09	97.09
	<questionmark>	97.37	92.50	94.87
	Overall	97.42	98.03	97.72
	ERR	0.37%		
	SER	4.4%		
asr.txt	<comma>	97.57	99.02	98.29
	<period>	98.26	97.05	97.65
	<questionmark>	97.97	92.16	94.98
	Overall	97.87	97.75	97.81
	ERR	0.38%		
	SER	4.0%		

4.4. Treinamento e Teste do Cenário 3

Em geral, na Tabela 5, não se revela melhorias significativas em relação ao Cenário 2. Contudo, a métrica de precisão na predição de vírgulas demonstra um desempenho geral levemente superior no *dataset* `asr.txt`, quando comparado com o `ref.txt`.

4.5. Treinamento e Teste do Cenário 4

A Figura 3 mostra um comportamento semelhante aos experimentos anteriores, com queda nas curvas de *loss* e *perplexity* ao longo do treinamento. No entanto, observa-

Tabela 5. Resultados da avaliação nos conjuntos de dados `ref.txt` e `asr.txt`.

Dataset	Punctuation	Precision	Recall	F1-score
ref.txt	<comma>	97.37	96.53	96.95
	<period>	97.70	98.32	98.01
	<questionmark>	95.12	90.70	92.86
	Overall	97.43	97.07	97.25
	ERR	0.44%		
	SER	5.2%		
asr.txt	<comma>	97.98	96.48	97.22
	<period>	98.11	97.07	97.59
	<questionmark>	97.18	96.38	96.77
	Overall	97.98	96.71	97.34
	ERR	0.47%		
	SER	4.7%		

se uma saturação mais rápida do modelo, o que contribui para resultados consistentes em termos de generalização.

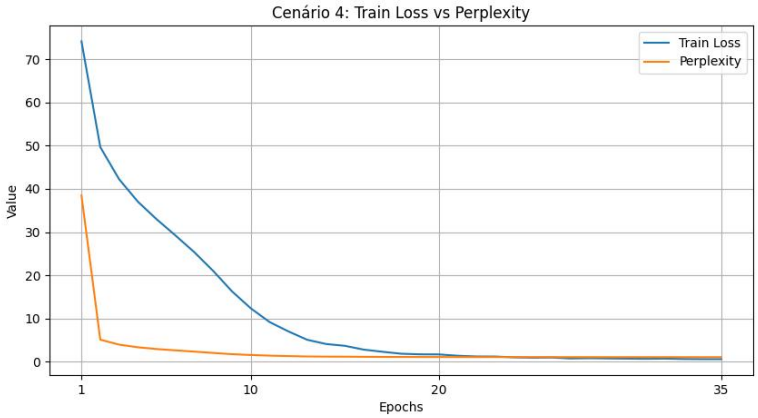

Figura 3. Gráfico de *Loss* e *Perplexity* do Cenário 4.

Os resultados apresentados na Tabela 6 revelam melhorias substanciais em todas as métricas em comparação com os cenários anteriores. A precisão, *recall* e *f1-score* atingiram seus valores máximos, indicando um desempenho superior na predição de pontuações. Além disso, os valores reduzidos de ERR e SER evidenciam uma significativa melhoria na capacidade do modelo de inserir pontuações de forma precisa e consistente nas transcrições.

4.6. Análise de Resultados

A avaliação do desempenho dos modelos treinados com os conjuntos de dados do IWSLT, distribuídos ao longo dos anos, evidencia uma tendência consistente de aprimoramento à medida que são utilizados dados mais recentes. Essa constatação indica que a evolução temporal dos corpora exerce influência direta e positiva na performance dos modelos avaliados.

Entre os experimentos realizados, o modelo treinado com a combinação dos dados referentes aos anos de 2014, 2015 e 2016 obteve os melhores resultados em todas as

Tabela 6. Resultados da avaliação nos conjuntos de dados `ref.txt` e `asr.txt`.

Dataset	Punctuation	Precision	Recall	F1-score
ref.txt	<comma>	99.59	97.93	98.76
	<period>	98.89	98.81	98.85
	<questionmark>	98.73	99.57	99.15
	Overall	99.25	98.39	98.82
	ERR	0.22%		
	SER	2.3%		
asr.txt	<comma>	99.64	97.95	98.76
	<period>	98.39	99.05	98.72
	<questionmark>	98.45	98.91	98.68
	Overall	99.05	98.46	98.75
	ERR	0.22%		
	SER	2.3%		

métricas analisadas. Esse resultado reforça a ideia de que tanto a diversidade temporal quanto o aumento no volume de dados contribuem de forma significativa para a robustez e eficiência da solução.

Cabe destacar que, mesmo diante de uma menor quantidade de amostras, os três primeiros cenários conseguiram prever pontos de interrogação. A união dos conjuntos proporcionou uma representação mais consistente dessa classe, o que possibilitou ao modelo um aprendizado mais eficaz. No entanto, o desafio de equilibrar as classes sem comprometer a integridade semântica dos dados permanece, evidenciando a necessidade de desenvolver estratégias inovadoras que favoreçam a representação semântica adequada durante o treinamento de modelos de linguagem.

5. Conclusão

Neste trabalho, teve-se como proposta desenvolver um modelo treinado de rede neural artificial, utilizando técnicas de processamento de linguagem natural, para a restauração de sentenças coloquiais no idioma português do Brasil, visando aprimorar a qualidade das transcrições de áudio. A metodologia empregada seguiu uma abordagem em duas etapas, envolvendo o treinamento e a inferência do modelo. O pré-processamento incluiu o mapeamento de símbolos de pontuação, a construção de vocabulário, o uso de *embeddings* GloVe de 300 dimensões e a utilização de uma rede *highway*.

A base de dados utilizada compreende os anos de 2014, 2015 e 2016 do IWSLT, com foco na restauração de pontuação em textos em português. Os experimentos foram organizados em quatro cenários: três com os anos individualmente e um com a combinação dos três. A arquitetura adotada inclui camadas Bi-LSTM, mecanismo de *self-attention* e decodificador CRF. A análise dos resultados apontou melhora contínua no desempenho com dados mais recentes, sendo a melhor performance alcançada com a junção dos três anos, destacando o impacto positivo da diversidade e do volume de dados.

Entre as limitações identificadas, destacam-se o tamanho reduzido da base e a dependência de um único corpus. Como possibilidades para trabalhos futuros, propõe-se: (i) a exploração de técnicas emergentes, como o Mini-BERT, com avaliação de seus requisitos computacionais; (ii) a análise comparativa com arquiteturas mais leves; (iii) a utilização de diferentes *embeddings*, como o Word2Vec; e (iv) a ampliação dos *datasets*, incorporando estilos e domínios variados para fortalecer a generalização do modelo.

Referências

Caseli, H. M. and Nunes, M. G. V., editors (2023). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN. `https://brasileiraspln.com/livro-pln`.

Chordia, V. (2021). Punktuator: A multilingual punctuation restoration system for spoken and written text. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 312–320. Association for Computational Linguistics.

de Lima, T. B., Rodrigues, L., Macario, V., Freitas, E., and Mello, R. F. (2023). Automatic punctuation verification of school students' essay in portuguese. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 58–70. SBC.

de Lima, T. B., Rolim, V., Nascimento, A. C., Miranda, P., Macario, V., Rodrigues, L., Freitas, E., Gašević, D., and Mello, R. F. (2024). Towards explainable automatic punctuation restoration for portuguese using transformers. *Expert Systems with Applications*, 257:125097.

Gris, L. R. S., Marcacini, R., Junior, A. C., Casanova, E., Soares, A., and Aluísio, S. M. (2023). Evaluating openai's whisper asr for punctuation prediction and topic modeling of life histories of the museum of the person.

Guerreiro, N. M., Rei, R., and Batista, F. (2021). Towards better subtitles: A multilingual approach for punctuation restoration of speech transcripts. In *Expert Systems with Applications*, volume 186, page 115740.

Lima, T. B. D., Miranda, P., Mello, R. F., Wenceslau, M., Bittencourt, I. I., Cordeiro, T. D., and José, J. (2022). Sequence labeling algorithms for punctuation restoration in brazilian portuguese texts. In *2022 11th Brazilian Conference(BRACIS)*, pages 616–630.

Moura, B. C. D., de S. Sales, A. G., de S. Linhares, J. E. B., Barbosa, F. M. D., and Neto, A. A. (2025). Avaliação in-domain e cross-domain em restauração de pontuação utilizando processamento de linguagem natural. *Anais do Computer on the Beach*, 16:45–52.

Olive, J., Christianson, C., and McCary, J. (2011). *Handbook of natural language processing and machine translation: DARPA global autonomous language exploitation*. Springer Science & Business Media.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP-2014)*, volume 12, pages 1532–1543.

Srivastava, R. K., Greff, K., and Schmidhuber, J. (2015). Highway networks. *arXiv preprint arXiv:1505.00387*.