PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Programa de Pós graduação em Informática

Diego Bernardes de Lima Santos

AVALIAÇÃO DA EFETIVIDADE DE *EMBEDDINGS* TEXTUAIS MULTILÍNGUES BASEADOS EM *TRANSFORMERS* PARA RECONHECIMENTO DE ENTIDADES NOMEADAS EM LÍNGUA PORTUGUESA

Belo Horizonte

2021

Diego Bernardes de Lima Santos

AVALIAÇÃO DA EFETIVIDADE DE *EMBEDDINGS* TEXTUAIS MULTILÍNGUES BASEADOS EM *TRANSFORMERS* PARA RECONHECIMENTO DE ENTIDADES NOMEADAS EM LÍNGUA PORTUGUESA

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para obtenção do título de Mestre em Informática.

Orientador: Prof. Dr. Wladmir Cardoso Brandão

Belo Horizonte

2021

FICHA CATALOGRÁFICA
Elaborada pela Biblioteca da Pontifícia Universidade Católica de Minas Gerais

Santos, Diego Bernardes de Lima
S237a Avaliação da efetividade de *embeddings* textuais multilíngues baseados em *transformers* para reconhecimento de entidades nomeadas em língua portuguesa / Diego Bernardes de Lima Santos. Belo Horizonte, 2021.
46 f. : il.

Orientador: Wladmir Cardoso Brandão
Dissertação (Mestrado) – Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática

1. Linguística - Processamento de dados. 2. Redes neurais (Computação). 3. Framework (Programa de computador). 4. Teoria da codificação. 5. Arquitetura de computador. 6. Transformadores elétricos. I. Brandão, Wladmir Cardoso. II. Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática. III. Título.

CDU: 681.3.091

Ficha catalográfica elaborada por Pollyanna Iara Miranda Lima - CRB 6/3320

Diego Bernardes de Lima Santos

AVALIAÇÃO DA EFETIVIDADE DE *EMBEDDINGS* TEXTUAIS MULTILÍNGUES BASEADOS EM *TRANSFORMERS* PARA RECONHECIMENTO DE ENTIDADES NOMEADAS EM LÍNGUA PORTUGUESA

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para obtenção do título de Mestre em Informática.

Prof. Dr. Wladmir Cardoso Brandão – PUC Minas

Prof. Dr. Humberto Torres Marques Neto – PUC Minas

Prof. Dr. Fernando Silva Parreiras – Universidade FUMEC

Dr. Frederico Giffoni de Carvalho Dutra – CEMIG

Prof. Dr. Evandrino Gomes Barros – CEFET-MG

Belo Horizonte, 08 de Abril de 2021.

Dedicatória

Este trabalho é dedicado às duas maiores mulheres que já conheci, que infelizmente não estão aqui para que eu possa abraçá-las: Minha avó Iracema e a minha mãe Lilian.

AGRADECIMENTOS

Gostaria de agradecer primeiramente a Deus, por ter sempre aberto as portas, me concedido saúde e força para enfrentar os desafios na minha vida.

Ao meu pai por toda a importância que sempre dedicou à minha formação.

À minha mãe por ter me ensinado os melhores valores que eu poderia ter, e mesmo não estando presente fisicamente, é muito presente em minha memória, no meu coração e nas minhas vitórias.

Meus irmãos Lorena e Rodrigo por terem compreendido meu momento difícil, me acolhido e torcido por mim em cada minuto dessa trajetória.

À Keise por ter me apoiado, incentivado e entendido que esse sonho era importante para mim, abdicando de muitos compromissos familiares, sendo compreensiva, e mesmo após nossa separação, sempre torceu para que eu conseguisse atingir esse objetivo.

Ao Professor Wladmir, que soube dos meus problemas pessoais, e de saúde, que aconteceram todos ao mesmo tempo, teve compreensão e principalmente paciência para me orientar nesse projeto.

À Thaísa, por todo o carinho, por ter me cedido ouvidos e ombros para chorar, mas principalmente por ter me convencido a não desistir, me ajudado a enxergar soluções, por ter me passado toda sua experiência acadêmica e ter sido uma pessoa central para que eu conseguisse me manter de pé nesse período.

"A persistência é o caminho do êxito."

Charles Chaplin

RESUMO

As abordagens recentes para a atividade de Reconhecimento de Entidades Nomeadas, uma das tarefas inerentes ao Processamento de Linguagem Natural, utilizam diversas arquiteturas, em sua maioria baseadas em Redes Neurais, para que seja possível desenvolver um aprendizado associado à linguagem, bem como executar a extração das entidades nomeadas, que geralmente são elementos relevantes dentro de textos ou fragmentos de textos. Normalmente quando um modelo de linguagem é treinado, utilizando um idioma específico, produz resultados efetivos, porém requer um volume maior de tempo e de recursos computacionais para a realização desse processo se comparado à utilização de um modelo multilingual pré-treinado, existente, que pode ser mais simples e rápido, se aplicados os ajustes finos necessários para o reconhecimento de entidades nomeadas. Este trabalho explora modelos multilinguais para reconhecimento de entidades nomeadas, adaptando e treinando arquiteturas baseadas em *transformers* para o português, que é uma linguagem complexa e desafiadora. Os experimentos mostram que as abordagens de embeddings de texto baseados em *transformers* multilinguais, com a aplicação de ajustes finos, utilizando um grande conjunto de dados superam os modelos de última geração, baseados em *transformers*, treinados especificamente para o português. Em particular, construímos um conjunto de dados abrangente de diferentes versões do HAREM para treinar nossa abordagem de embeddings de texto baseada em *transformers* multilinguais, que obteve 88,0% de precisão e 87,8% em F1 no reconhecimento de entidades nomeadas para o português, com ganhos de até 9,89% de precisão e 11,60% na F1 em comparação com o estado da arte da abordagem treinada especificamente para o português.

Palavras-chave: Reconhecimento de Entidades Nomeadas, Embeddings de Texto, Redes Neurais, Transformer, Multilingual, Português.

ABSTRACT

Recent state of the art named entity recognition approaches are based on deep neural networks that use an attention mechanism to learn how to perform the extraction of named entities from relevant fragments of text. Usually, training models in a specific language leads to effective recognition, but it requires a lot of time and computational resources. However, fine-tuning a pre-trained multilingual model can be simpler and faster, but there is a question about how effective that recognition model can be. This article exploits multilingual models for named entity recognition by adapting and training tranformer-based architectures for Portuguese, a challenging complex language. Experimental results show that multilingual trasformer-based text embeddings approaches fine tuned with a large dataset outperforms state of the art trasformer-based models trained specifically for Portuguese. In particular, we build a comprehensive dataset from different versions of HAREM to train our multilingual transformer-based text embedding approach, which achieves 88.0% of precision and 87.8% in F1 in named entity recognition for Portuguese, with gains of up to 9.89% of precision and 11.60% in F1 compared to the state of the art single-lingual approach trained specifically for Portuguese.

Keywords: Named Entity Recognition, Text Embedding, Neural Network, Transformer, Multilingual, Portuguese.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

alBERT *A Lite BERT*

BART *Bidirectional and Auto-Regressive Transformers*

BERT *Bidirectional Encoder Representations from Transformers*

BiLSTM *Bidirectional long short-term memory*

CRF *Conditional Random Fields*

LG *Local Grammar*

LSTM *Long short-term memory*

ML Modelos Multilíngues

NER Named Entity Recognition

NLP Natural Language Processing

NSP *Nex Sentence Prediction*

PT Português

RN Redes Neurais Profundas

roBERTa *Robustly Optimized BERT Approach*

SOTA *State-of-The-Art*

tsv *Tab-separated values*

xml *Extensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

Processamento de linguagem natural, do inglês Natural Language Processing (NLP) é um campo de pesquisa da ciência da computação com várias aplicações, tais como leitura automática de texto, sistemas de respostas a perguntas, interpretação de conteúdo de áudio, classificação de documentos e análise preditiva de fragmentos de texto. Normalmente, os sistemas de NLP executam um conjunto de tarefas básicas de pré-processamento no texto de entrada, como análise, tokenização, remoção de palavras irrelevantes (*stopwords*), lematização e marcação.

Especificamente, a tarefa de Reconhecimento de Entidades Nomeadas, do inglês Named Entity Recognition (NER) é uma tarefa de marcação de NLP, que extrai informações semanticamente importantes, identificando-as no texto, por exemplo, nomes de pessoas, lugares, valores monetários e temporais (BORTHWICK, 1999). Os elementos extraídos são entidades relevantes para a compreensão do conteúdo textual, que fazem sentido dentro de um contexto. Por exemplo, o reconhecimento da entidade "New York" como um local em uma frase pode ser importante para detectar onde um determinado evento ocorreu ou mesmo para relacionar esse local a outros locais, manipulando e interpretando entidades semelhantes ou com o mesmo valor semântico.

1.1 Problema

NER é fortemente dependente do contexto, ou seja, palavras ou expressões podem ser reconhecidas como diferentes tipos de entidade em diferentes contextos. Por exemplo, na frase "Maria reza a São Paulo pela saúde", a expressão "São Paulo"se refere a uma pessoa (entidade religiosa), mas na frase "Nos mudaremos para São Paulo no próximo ano", a expressão "São Paulo"refere-se a um lugar (entidade de localização). Mesmo que a grafia de uma palavra ou expressão citada em frases diferentes seja idêntica, o significado pode ser distinto em contextos diferentes. Além disso, as frases são formuladas em diferentes maneiras em diferentes idiomas, e os idiomas diferem uns dos outros em estrutura, forma e complexidade, o que impõe questões ainda mais desafiadoras para NER.

Abordagens tradicionais de NER usam estratégias baseadas em gramática linguística ou modelos estatísticos que requerem grande quantidade de dados de treinamento anotados manualmente para reconhecer entidades no texto, através da aplica-

ção de regras inferidas ou descritas pelo modelo desenvolvido (MARSH; EPERZA-NOWSKI, 1998).

Por anos, *Conditional Random Fields* (CRF) tem sido a estratégia de última geração para NER, considerando o contexto em um modelo de aprendizagem que suporta dependências sequenciais entre predições (LAFFERTY; MCCALLUM; PEREIRA, 2001).

Recentemente, abordagens baseadas em redes neurais profundas alcançaram resultados ainda mais eficazes do que o CRF para NER (GOLDBERG, 2016). Eles aprendem representações de texto distribuídas (embeddings de texto) a partir de uma grande quantidade de texto para construir um modelo de linguagem que pode ser usado com eficácia em várias tarefas de NLP, incluindo NER.

Modelos neurais profundos monolíngues (modelos de NLP de treinamento em um idioma específico) geralmente levam ao reconhecimento efetivo de entidades, exigindo, porém, muito tempo e recursos computacionais para o treinamento. Além disso, tais abordagens monolíngues requerem uma grande quantidade de dados em cada idioma específico para o treinamento, eventualmente não disponíveis ou de obtenção restritiva para determinados idiomas. No entanto, o ajuste fino de um modelo multilíngue pré-treinado pode ser mais barato, mais simples e rápido, não exigindo nenhum conjunto de dados de treinamento específico em um único idioma e menos tempo e recursos computacionais para treinamento. Mas, como modelos NER multilíngues eficazes podem ser comparados aos modelos monolíngues, especialmente para idiomas complexos, como o português?

1.2 Objetivos

Nas seções a seguir, serão tratados os objetivos deste trabalho, considerando a avaliação de implementações de reconhecimento de entidades para o idioma português e como combiná-las ou adaptá-las para contribuir com a melhoria dos resultados para este idioma.

1.2.1 Objetivos Gerais

Neste artigo, temos como objetivo explorar modelos multilíngues para NER, adaptar e treinar *embeddings* de texto baseados em *transformers* para reconhecimento de entidades nomeadas em português. Particularmente, propor uma abordagem NER para o português, treinar e ajustar um modelo de NLP multilíngue baseado em *transformers*, usando um conjunto de dados abrangente reúne e integra diferentes versões do conjunto de dados HAREM (SANTOS; CARDOSO, 2007). Além disso, avaliar nossa

abordagem proposta comparando-a com a abordagem monolíngue do estado da arte, do inglês, *State-of-The-Art* (SOTA) para NER em português.

1.2.2 Objetivos Específicos

Os objetivos específicos deste trabalho envolvem os seguintes aspectos para análise das abordagens de NER em português:

- Elencar abordagens de NER com melhor acurácia média entre diversos idiomas.
- Adaptar as abordagens de NER para o idioma português realizando o treinamento com conjuntos de dados relevante e adequado para o treinamento.
- Propor ajustes finos às abordagens em *Transformers* para que o processamento em *embeddings* de texto atinja resultados relevantes.
- Estabelecer análise comparativa das abordagens analisadas.
- Propor abordagem que consiga obter níveis de acurácia superiores aos trabalhos no estado da arte para o idioma português.

1.3 Justificativa

O mercado de aplicações que utilizam processamento natural cada vez mais abarca maiores possibilidades na utilização dessas abordagens para o crescimento de seus negócios. A análise de sentimentos dos clientes, os e-mails e comunicações com atendimento automático, os chatbots que fazem desde atendimento, até o processo de vendas, bem como análises de conteúdo estratégico, fazem crescer a demanda por abordagens de NLP, mas também, a tarefa de NER que é relevante na obtenção de informações centrais de um conteúdo, até mesmo para resumo ou filtragem do que é relevante ao usuário.

Nesse contexto, o trabalho aqui proposto traz relevante contribuição no sentido de proporcionar ao público em geral, usuário da língua portuguesa, uma forma de aplicar, com resultados importantes, as tarefas de NER exemplificadas no exemplo anterior, outras além dessas, e especificamente para a língua portuguesa.

Resultados experimentais mostram que nossa abordagem multilíngue para NER em português supera a abordagem monolíngue SOTA com ganhos de até 9,89% de precisão e 11,60% em F1, alcançando 88,0% de precisão e 87,8% em F1 no reconhecimento de entidades nomeadas para o português.

As principais contribuições deste trabalho são:

- Proposição de um conjunto de dados abrangente para aprimorar o treinamento de modelos NER para o português combinando diferentes versões do conjunto de dados HAREM.
- Proposição de abordagem NER multilíngue para o português, ajustando e treinando diferentes redes neurais baseadas em *transformers* para NER multilíngue.
- Fornecemos uma avaliação completa de nossa abordagem proposta, contrastando-a com a abordagem SOTA monolíngue para NER em português relatada na literatura.

1.4 Organização da Dissertação

Este trabalho está organizado da seguinte maneira: a seção 2 apresenta os fundamentos teóricos do reconhecimento de entidades nomeadas, *embeddings* de palavras e arquiteturas de redes neurais baseadas em *transformers*. A seção 3 apresenta trabalhos relacionados relatados na literatura para NER, incluindo o estado da arte da abordagem para NER em português. A seção 4 apresenta nossa abordagem NER multilíngue para o português, bem como o conjunto de dados abrangente que criamos para melhorar o treinamento de nossa abordagem. A seção 5 apresenta a descrição e os resultados dos experimentos que realizamos para avaliar a abordagem proposta. Finalmente, a Seção 6 apresenta as conclusões deste trabalho, apontando direções para trabalhos futuros.

2 REFERENCIAL TEÓRICO

NER é uma tarefa de NLP que identifica pessoas, localização, moeda e outras informações relevantes em um texto (BORTHWICK, 1999). Enquanto as abordagens NER tradicionais usam estratégias baseadas em gramática linguística ou modelos estatísticos que requerem uma grande quantidade de dados de treinamento anotados manualmente para reconhecer entidades no texto (MARSH; EPERZANOWSKI, 1998), as abordagens NER recentes usam redes neurais profundas para aprender um reconhecimento eficaz modelo (GOLDBERG, 2016). Em particular, eles aprendem *embeddings* de texto de uma grande quantidade de texto para construir um modelo de linguagem que pode ser usado com eficácia para NER.

2.1 *Word Embeddings*

Recentemente, surgiram diferentes formas de representar texto, permitindo análises mais precisas de informações textuais, por exemplo, a análise de semelhança entre duas palavras. Uma representação de texto distribuída, ou *embeddings* de texto, pode ser gerada por abordagens de Redes Neurais Profundas (RN) que aprendem modelos de linguagem a partir de uma grande quantidade de corpus de linguagem natural. Em particular, *embeddings* de palavras são representações vetoriais contínuas que descrevem o significado dos termos (LEVY; GOLDBERG, 2014). Normalmente, essa representação distribuída é um vetor de valor real contínuo não mutuamente exclusivo e de comprimento fixo aprendido por um RN, normalmente muito menor que o tamanho do vocabulário (BENGIO et al., 2003).

As representações de vetores contínuos são capazes de representar palavras sintaticamente, mas também permitem o aprendizado de valores semânticos de termos, ou seja, *embeddings* de palavras podem capturar semelhanças entre palavras com significado semelhante, mesmo que a grafia seja bastante diferente entre elas (MIKOLOV et al., 2013b) A Figura 1 apresenta grupos de palavras com contexto semelhante medido pela similaridade de cosseno entre *embeddings* de palavras.

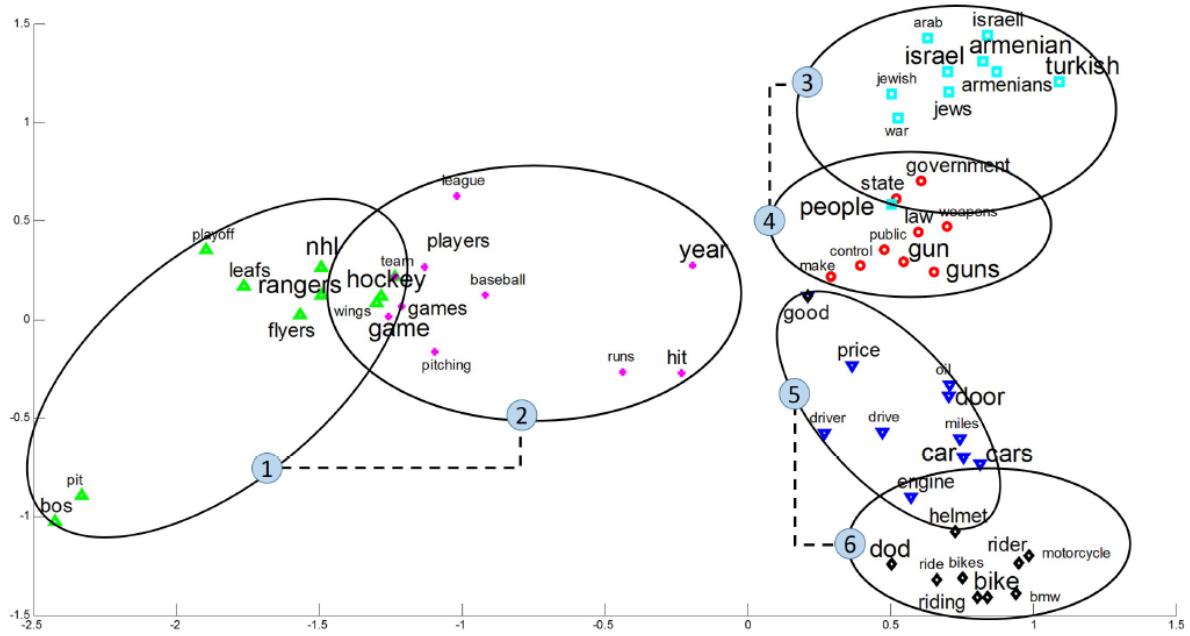

Figura 1 – Correlação entre palavras representadas como vetores contínuos medidos pela similaridade de cosseno.
Fonte: (XUN et al., 2017)

Nos últimos anos, diferentes *frameworks* e algoritmos para geração de *embeddings* de palavras têm sido propostos, particularmente WORD2VEC, GLOVE, e FASTTEXT. WORD2VEC (MIKOLOV et al., 2013a, 2013b) é um framework composto pelos primeiros modelos de incorporação de palavras eficientes, particularmente o BoW contínuo (CBOW) e o skip-gram contínuo (SKIP-GRAM), para aprender representações distribuídas de palavras de grande quantidade de texto não estruturado com bilhões de palavras. O treinamento de tais modelos não requer multiplicações de matrizes densas e pode ser feito em um dia em um conjunto de dados de cem bilhões de palavras com uma única máquina.

Particularmente, o modelo CBOW é uma simplificação da primeira abordagem de modelo de linguagem neural prática proposta na literatura (BENGIO et al., 2003) que usa uma rede neural feedforward totalmente conectada para aprender simultaneamente uma representação distribuída de palavras e a função de distribuição de probabilidade conjunta para essas representações de palavras a partir de um enorme corpus de texto em linguagem natural com milhões de palavras.

No modelo CBOW a camada oculta não linear é removida, a camada de projeção é compartilhada por todas as palavras e o contexto da palavra é capturado por um classificador log-linear treinado para prever uma palavra-alvo dados seus dois anteriores e dois próximos palavras vizinhas. O modelo SKIP-GRAM é semelhante a CBOW, mas o classificador log-linear é treinado para prever as duas palavras vizinhas ante-

riores e as duas próximas dadas uma palavra-alvo (MIKOLOV et al., 2013a). Além disso, os modelos realizam subamostragem de palavras frequentes, resultando em um treinamento mais rápido e representações aprimoradas de palavras incomuns. Além disso, eles usam dois métodos de treinamento de substituição para softmax completo, resultando em aceleração e representações distribuídas precisas, especialmente para palavras frequentes (MIKOLOV et al., 2013b). Os métodos de treinamento de substituição são softmax hierárquico (MORIN; BENGIO, 2005) e amostragem negativa, um NCE simplificado (GUTMANN; HYVÄRINEN, 2012).

Uma propriedade interessante da palavra *embeddings* aprendida pelos modelos Word2Vec é que operações simples de vetor podem frequentemente produzir resultados significativos. Por exemplo, a operação de soma entre o vetor (*USA*) e o vetor (*capital*) resulta em um vetor próximo ao vetor (*Washington*). Além disso, os *embeddings* de palavras podem ser combinados usando operações simples para representar trechos de texto mais longos, como frases, parágrafos e documentos. Por exemplo, o vetor (*Boston*) e o vetor (*Globo*) podem ser combinados para obter o vetor (*Globo de Boston*). No entanto, a palavra *embedding* resultante geralmente é incapaz de representar frases idiomáticas que não sejam composições de palavras individuais, como *globo de Boston*. Além disso, os *embeddings* de palavras aprendidos pelos modelos Word2Vec exibem uma estrutura linear que torna possível o raciocínio analógico preciso. Por exemplo, o vetor (*queen*) sendo a representação mais próxima do vetor (*king*) menos o vetor (*man*) mais o vetor (*mulher*) fornecem uma maneira de teste o par de analogia *homem*:*king* :: *mulher*:*queen*.

GloVe incorpora estatísticas globais de ocorrências de palavras tipicamente capturadas por modelos de linguagem baseados em contagem em um modelo log-bilinear para aprendizagem não supervisionada de embeddings de palavras (PENNINGTON; SOCHER; MANNING, 2014). A intuição é que os modelos de *shallow window*, como skip-gram, utilizam mal as estatísticas do corpus, pois treinam na janela de contexto local em vez de em contagens de co-ocorrência global. Portanto, treinar um modelo RN simultaneamente no contexto local e em contagens globais de coocorrência palavra-palavra, fazendo uso eficiente de estatísticas, produz embeddings de palavras com subestrutura significativa.

FastText é outra abordagem simples e não supervisionada que aprende representações distribuídas considerando unidades de sub-palavra e representando palavras por uma soma de seus caracteres n-gramas (BOJANOWSKI et al., 2017). É uma extensão do modelo de skip-gram contínuo (KIROS et al., 2015) que incorpora n-gramas, levando em consideração a estrutura interna das palavras, o que é importante para idiomas morfologicamente ricos, onde muitas formações de palavras seguem regras. Por exemplo, nas línguas latinas, a maioria dos verbos tem mais de dezenas de formas

flexionadas diferentes. Esses idiomas contêm muitas formas de palavras que ocorrem raramente (ou nunca ocorrem) no corpus de treinamento, tornando difícil aprender boas representações de palavras. Além disso, FastText é capaz de construir vetores de palavras para palavras que não aparecem no conjunto de treinamento. Resultados experimentais em tarefas de similaridade e analogia de palavras mostram que FastText supera modelos de Word2Vec que não levam em consideração informações de sub-palavra, bem como métodos baseados em análise morfológica em nove idiomas diferentes.

2.2 Transformers

Os Transformers são modelos de transdução de sequência baseados exclusivamente na atenção, substituindo as camadas recorrentes mais comumente usadas em arquiteturas codificador-decodificador por *multi-headed self-attention*, consequentemente permitindo incremento da paralelização (VASWANI et al., 2017). Em particular, segue uma estrutura de codificador-decodificador usando auto-atenção empilhada e camadas totalmente conectadas para o codificador e decodificador, onde o codificador mapeia uma sequência de entrada de representações de símbolos para uma sequência de representações contínuas, alimentando o decodificador que gera uma sequência de saída de símbolos, um elemento por vez. Em cada etapa, o modelo é auto-regressivo, consumindo os símbolos gerados anteriormente como entrada adicional ao gerar a próxima. Resultados experimentais em tradução automática e análise de conteúdo em inglês mostram que os Transformers superam os modelos discriminativos de referência (*baseline*) por uma fração do custo de treinamento.

O mecanismo de atenção é uma forte diferenciação entre Transformers e outras arquiteturas RN, permitindo estimar a correlação entre os elementos de forma bidirecional. Normalmente, existem dois mecanismos de atenção:

- Auto-atenção: intra-análise dos vetores de *embeddings* de uma frase, realizando o cálculo de similaridade entre diferentes palavras dentro de uma mesma frase. Nesta análise, o mecanismo extrai a correlação entre as palavras da frase. O sentido dos vetores representa se as palavras têm valores semânticos semelhantes ou distintos.

- Atenção *multi-head*: divide as sentenças em partes menores para realizar o cálculo de similaridade entre as matrizes. É semelhante ao mecanismo de auto-atenção, mas entre diferentes partes das frases, identificando a relação entre as palavras por meio de segmentos de texto, chamados de subespaços.

A Figura 2 apresenta o mecanismo de atenção que estima a correlação entre palavras com valores semânticos semelhantes, de forma bidirecional. Ainda na Figura 2 podemos observar que a palavra "making"tem uma relação próxima com as palavras "2009"e "laws", por exemplo, a palavra "making"aparece nas mesmas expressões que "2009"e "laws". Essa relação permite a previsão dos próximos termos em frases com palavras com significados semelhantes.

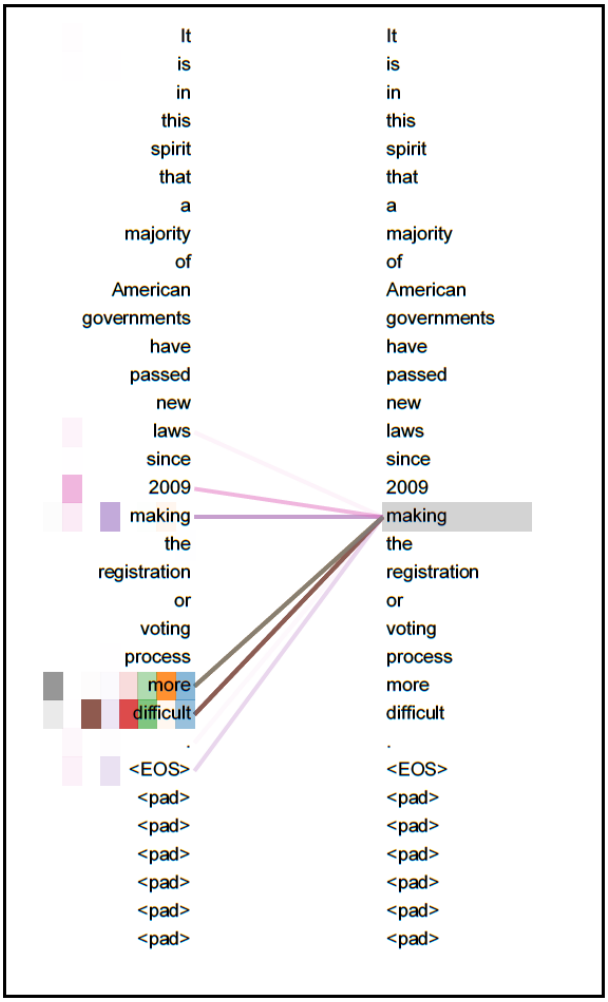

Figura 2 – Correlação de Palavras - Mecanismo de Atenção. Fonte: (VASWANI et al., 2017)

2.2.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) é uma abordagem de representação de linguagem projetada para pré-treinar representações bidirecionais profundas de texto não rotulado, condicionando conjuntamente os contextos esquerdo e direito em todas as camadas (DEVLIN et al., 2018). Em particular, um TRANSFORMER bidirecional profundo é pré-treinado em um modelo de linguagem mascarada e objeti-

vos de previsão da próxima frase, permitindo que a representação concatene o contexto esquerdo e direito, reduzindo assim a necessidade de muitas tarefas específicas e engenharia complexa dentro das abordagens. BERT é o primeiro modelo de representação baseado em ajuste fino que alcança desempenho de ponta em um grande conjunto de tarefas em nível de frase e *token*, superando muitas arquiteturas de tarefas específicas.

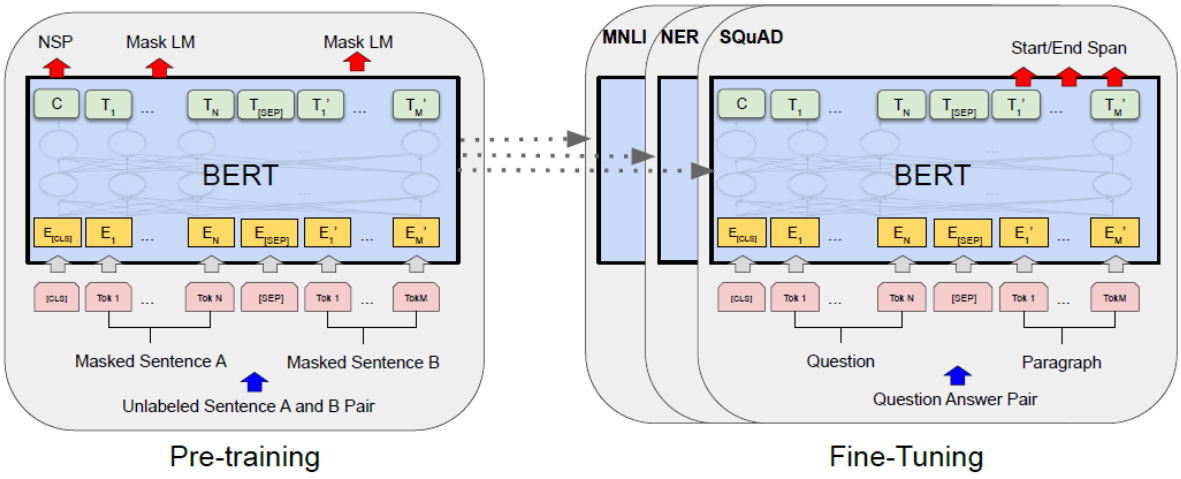

Figura 3 – Camadas do BERT.
Fonte: (DEVLIN et al., 2018)

A Figura 3 apresenta as camadas RN da arquitetura BERT. Particularmente, podemos observar as etapas de pré-treinamento e de ajuste fino. Durante o pré-treinamento, o conjunto de dados de entrada é usado sem rótulos, realizando assim o treinamento não supervisionado dos dados. Existem duas tarefas principais durante esta fase:

- Mascaramento de *tokens*: selecionando aleatoriamente uma porcentagem de cerca de 15% dos *tokens* de entrada e aplicando uma máscara sobre eles para que o treinamento faça a predição desses *tokens*.
- *Nex Sentence Prediction* (NSP): treino para tarefa de resposta a perguntas, prevendo quais são as sentenças subsequentes em relação às anteriores.

Após a etapa de pré-treinamento, os dados de saída podem ser usados como entrada para outras tarefas de NLP. Na Figura 3, podemos observar que os dados de saída do pré-treinamento são usados para tarefas de inferência de linguagem natural (MNLI), de reconhecimento de entidade nomeada (NER) e de resposta a perguntas (SQuAD).

2.2.2 RoBERTa

Robustly Optimized BERT Approach (roBERTa) é um *framework* baseado em BERT para pré-treinamento de modelo de linguagem que estende BERT treinando o modelo com *batches* maiores, maior volume de dados, e em sequências mais longas, também removendo a tarefa de previsão da próxima frase e alterando dinamicamente o padrão de mascaramento aplicado aos dados de treinamento (Liu et al., 2019). Resultados experimentais em tarefas NLP usando os benchmarks GLUE, RACE e SQuAD mostram que RoBERTa atinge resultados no estado da arte, superando BERT e XLNet, uma abordagem de aprendizagem autorregressiva (YANG et al., 2019). A tabela 1 apresenta os parâmetros experimentais e os resultados (desempenho medido por precisão) comparando RoBERTa e BERT em três tarefas diferentes: resposta a perguntas (SQuAD), inferência em linguagem natural (MNLI) e classificação de sentenças (SST).

	BERT-Large	RoBERTa
Data	13GB	16GB
Batches	256	8K
Steps	1M	100K
SQuAD v1.1	90,9	93,6
SQuAD v2.0	81,8	87,3
MNLI-m	86,6	89.0
SST-2	93,7	95.3

Tabela 1 – BERT/RoBERTa parâmetros e desempenho.
Fonte: (Liu et al., 2019)

De acordo com o que podemos observar na Tabela 1, existem diferenças significativas no treinamento, com alterações no tamanho dos *batches* e no número de etapas do treinamento. Enquanto RoBERTa usa um conjunto de dados maior que o BERT para realizar seu treinamento, *batches* vigorosamente maiores para seu processamento, porém o processamento ocorre em um número menor de etapas. RoBERTa é uma abordagem robusta, porém, como pode ser visto nas tarefas SQuAD, MNLI e SST, RoBERTa apresenta resultados semelhantes e até melhores do que na abordagem BERT.

2.2.3 DistilBERT

DistilBERT (destilado BERT) é uma versão pré-treinada menor e mais rápida de propósito geral do BERT, que retém quase as mesmas capacidades de compreensão da linguagem (SANH et al., 2019). Em particular, ele usa modelos de linguagem pré-treinados com destilação de conhecimento (*distilation*), uma técnica de compressão em que um modelo compacto é treinado para reproduzir o comportamento de um modelo

maior ou um conjunto de modelos, resultando em modelos mais leves e rápidos no tempo de inferência, enquanto também requer menor treinamento computacional. Particularmente, o resultado mantém 97% de compreensão da linguagem em seu modelo com aproximadamente 60% de redução no tamanho do modelo, rodando em torno de 60% mais rápido. DistilBERT pode ser ajustado em várias tarefas NLP, mantendo a flexibilidade de modelos maiores enquanto é pequeno o suficiente para ser executado na ponta, por ex. em dispositivos móveis.

A técnica de destilação (HINTON; VINYALS; DEAN, 2015) consiste em treinar um modelo, baseado em um modelo maior, denominado professor, que é usado para ensinar o modelo destilado, denominado aluno, a reproduzir o comportamento do modelo maior. Assim, DistilBERT é um modelo mais enxuto baseado no comportamento do modelo BERT original. A tabela 2 apresenta uma comparação de desempenho (precisão) em diferentes tarefas de NLP entre BERT, DistilBERT e ELMo, uma abordagem de representação de palavras contextualizada profunda que modela características sintáticas e semânticas complexas do uso das palavras e como esses usos variam entre os contextos linguísticos (polissemia) (PETERS et al., 2018).

	Score	CoLA	MNLI	QNLI
ELMo	68.7	44.1	68.6	76.6
BERT-Base	79.5	56.3	86.7	88.6
DistilBERT	77.0	51.3	82.2	87.5

Tabela 2 – Desempenho de BERT, DistilBERT e ELMo.

Conforme apresentado na Tabela 2 observamos que o desempenho do DistilBERT é próximo ao do BERT, mesmo atuando com um modelo reduzido.

2.2.4 ALBERT

A Lite BERT (alBERT) é outra arquitetura eficiente baseada em BERT com significativamente menos parâmetros do que uma arquitetura BERT tradicional (LAN et al., 2019). Em particular, ALBERT incorpora técnicas de redução de parâmetros que levantam os principais obstáculos no dimensionamento de modelos pré-treinados, atuando também como uma forma de regularização que estabiliza o treinamento e auxilia na generalização. Primeiro, ele incorpora a parametrização de incorporação fatorada, ou seja, decompor a matriz de incorporação de vocabulário grande em duas matrizes pequenas, separando assim o tamanho das camadas ocultas do tamanho de incorporação de vocabulário, tornando mais fácil aumentar o tamanho oculto sem aumentar significativamente o tamanho do parâmetro dos *embeddings* de vocabulário. Em segundo lugar, incorpora o compartilhamento de parâmetros entre camadas, evitando que o pa-

râmetro cresça com a profundidade da rede. Além disso, ALBERT substitui a próxima previsão de frase proposta no BERT original por uma perda auto-supervisionada para a previsão de ordem de frase. Experimentos com benchmarks GLUE, RACE e SQuAD mostram que ALBERT alcança desempenho de ponta em tarefas de compreensão de linguagem natural superando BERT, XLNet e RoBERTa.

Em particular, ALBERT aborda o problema de escalabilidade de BERT derivado de questões de consumo de memória. O crescimento do número de parâmetros do BERT tornou-se um desafio importante devido ao alto consumo de memória. Poucos trabalhos relatados na literatura abordam este problema, usando paralelismo (SHAZEER et al., 2018) ou gerenciando efetivamente o consumo de memória através de um mecanismo de limpeza para minimizar o impacto no desempenho (Gomez et al., 2017). No entanto, o obstáculo criado pelo *overhead* de comunicação da arquitetura BERT não é abordado por esses trabalhos relatados.

Assim, o BERT foi estendido por ALBERT a fim de reduzir em torno de 89% o número de parâmetros, melhorando o desempenho em tarefas de NLP. A tabela 3 mostra uma comparação entre os hiper-parâmetros de BERT e ALBERT.

Model	Parameters		Layer		Embedding Size
	#	Sharing	#	Hidden	
BERT-Base	108M	No	12	768	768
BERT-Large	334M	No	24	1024	1024
ALBERT-Base	12M	Yes	12	128	768
ALBERT-Large	18M	Yes	24	128	1024
ALBERT-XLarge	60M	Yes	24	128	2048
ALBERT-XXLarge	235M	Yes	12	128	4096

Tabela 3 – Hiper-parâmetros de BERT e ALBERT.
Fonte: (LAN et al., 2019)

Mesmo usando menos hiper-parâmetros do que BERT, ALBERT fornece resultados aprimorados em diferentes tarefas de NLP, como SQuAD v1.1 (+1.9%), SQuAD v2.0 (+3.1%), MNLI (+1.4%), SST-2 (+2,2%) e RACE (+8,4%) usando relativamente menos recursos e com uma fase de treinamento mais rápida (LAN et al., 2019).

2.2.5 BART

Bidirectional and Auto-Regressive Transformers (BART), é uma abordagem baseada em BERT, cuja etapa de treinamento é realizada por documentos que são inicialmente corrompidos par que em seguida seu conteúdo seja reconstruído pela aplicação minimizando as perdas de dados no processo de reconstrução. O que diferencia BART de

outras abordagens é a flexibilidade de inclusão de ruído de modo a permitir a generalização do mascaramento dos termos originais, forçando o modelo a compreender o conteúdo geral da frase e transformá-la novamente para que seja o mais similar possível à entrada.

	SQuAD 1.1	SQuAD 2.0	MNLI	SST	QQP	QNLI	STS-B
BERT	**84.1/90.9**	**79.0/81.8**	**86.6/-**	**93.2**	**91.3**	**92.3**	**90.0**
UniLM	-/-	80.5/83.4	87.0/85.9	94.5	-	92.7	-
XLNet	89.0/94.5	86.1/88.8	89.8/-	95.6	91.8	93.9	91.8
RoBERTa	**88.9/94.6**	**86.5/89.4**	**90.2/90.2**	**96.4**	**92.2**	**94.7**	**92.4**
BART	**88.8/94.6**	**86.1/89.2**	**89.9/90.1**	**96.6**	**92.5**	**94.9**	**91.2**

Tabela 4 – Comparação dos Resultados em tarefas NLP - Bart x BERT
Fonte: (LEWIS et al., 2019)

A Tabela 4 apresenta alguns resultados de aplicações que utilizaram BERT, roBERTa e BART para tarefas de NLP. BART possui resultados interessantes, que foram muito próximos dos resultados de roBERTa, por exemplo, que também é um modelo com resultados expressivos, além disso, o BART superou os resultados de BERT em todas as tarefas NLP observadas no experimento.

3 TRABALHOS RELACIONADOS

O surgimento de abordagens que usam *Transformers* para melhorar o desempenho em tarefas de NLP tem crescido consideravelmente nos últimos anos. Particularmente para NER em línguas complexas, um trabalho recente relatado na literatura (ARKHIPOV et al., 2019) usa *Transformers* para o reconhecimento de entidades nomeadas em línguas eslavas, alcançando até 93% de desempenho na medida F1 quando aplicado à língua tcheca.

Recentemente, diferentes arquiteturas RN foram propostas para realizar NER em português (SOUZA; NOGUEIRA; LOTUFO, 2019). Além da análise comparativa entre as arquiteturas, os autores propuseram uma abordagem eficaz tanto para a geração de *embbedings* de palavras quanto para o reconhecimento de entidades nomeadas em português. A abordagem proposta usa BERT para, inicialmente, gerar a palavra em *embedding* para o português e, por fim, usar esse *embeddings* para NER. Os autores também avaliam diferentes arquiteturas RN, como *Long short-term memory* (LSTM) e *Bidirectional long short-term memory* (BiLSTM) para o reconhecimento de entidades nomeadas em português. Os autores também combinam essas diferentes arquiteturas com CRF (LAFFERTY; MCCALLUM; PEREIRA, 2001) para melhorar o desempenho.

Na Figura 4 é possível observar o desenho da abordagem proposta por (SOUZA; NOGUEIRA; LOTUFO, 2019), considerando o fluxo a partir dos dados de entrada, que são tokenizados e codificados para os *embeddings* de texto no padrão BERT de forma semelhante a abordagem proposta por esse trabalho, porém, na camada de saída, uma camada adicional utilizando CRF para que as entidades sejam identificadas e extraídas, apresentando na saída a classificação dos termos.

Ainda na Figura 4, pode-se indicar a diferença do módulo BERT em relação a este trabalho, na proposta de (SOUZA; NOGUEIRA; LOTUFO, 2019), a camada BERT é responsável pela codificação dos tokens de entrada em *embeddings* de texto, pontuação das categorias em relação às expressões, porém o módulo seguinte, o CRF, através de seu modelo probabilístico realiza a junção dos termos com as respectivas predições de entidades nomeadas.

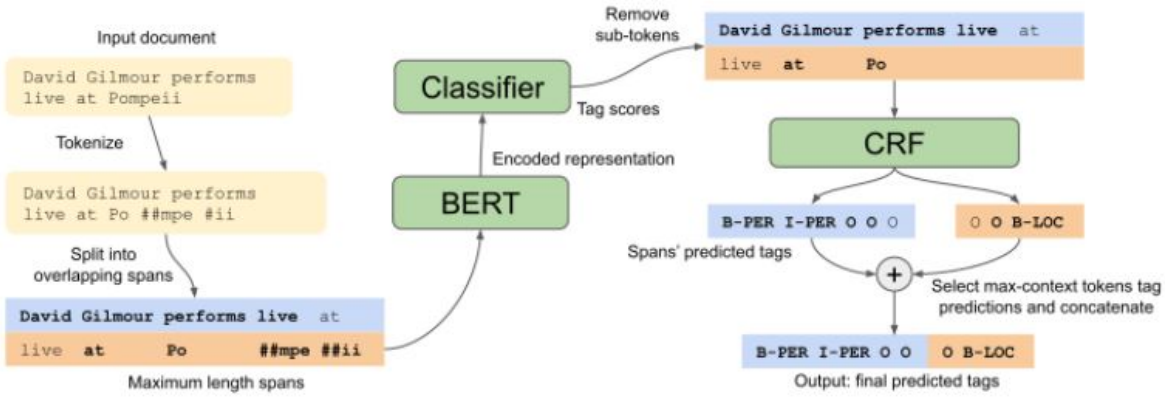

Figura 4 – Abordagem de NER em Português utilizando BERT e CRF. Fonte: (SOUZA; NOGUEIRA; LOTUFO, 2019)

A Tabela 5 resume os resultados experimentais das arquiteturas propostas para Modelos Multilíngues (ML) e Português (PT) em dois cenários: um cenário completo usando todo o conjunto de dados HAREM com 10 classes, e outro cenário seletivo usando um subconjunto de 5 classes de HAREM onde a abordagem proposta tem melhor desempenho.

Abordagem	Cenário Completo			Cenário Seletivo		
	Prec.	Revoc.	F1	Prec.	Revoc.	F1
CharWNN	67.16	63.74	65.41	73.98	68.68	71.23
LSTM-CRF	72.78	68.03	70.33	78.26	74.39	76.27
BiLSTM-CRF+FlairBBP	74.91	74.37	74.64	83.38	81.17	82.26
ML-BERT-Base	2.97	73.78	73.37	77.35	79.16	78.25
ML-BERT-Base-CRF	74.82	73.49	74.15	80.10	78.78	79.44
ML-BERT-Base-LSTM	69.68	69.51	69.59	75.59	77.13	76.35
ML-BERT-Base-LSTM-CRF	74.70	69.74	72.14	80.66	75.06	77.76
PT-BERT-Base	78.36	**77.62**	77.98	83.22	82.85	83.03
PT-BERT-Base-CRF	78.60	76.89	77.73	83.89	81.50	82.68
PT-BERT-Base-LSTM	75.00	73.61	74.30	79.88	80.29	80.09
PT-BERT-Base-LSTM-CRF	78.33	73.23	75.69	84.58	78.72	81.66
PT-BERT-Large	78.45	77.40	77.92	83.45	**83.15**	**83.30**
PT-BERT-Large-CRF	**80.08**	77.31	**78.67**	**84.82**	81.72	83.24
PT-BERT-Large-LSTM	72.96	72.05	72.50	78.13	78.93	78.53
PT-BERT-Large-LSTM-CRF	77.45	72.43	74.86	83.08	77.83	80.37

Tabela 5 – Desempenho em precisão, revocação e F1 da abordagem monolíngue SOTA para NER treinada especificamente para o português em dois cenários experimentais.

Na Tabela 5 observamos que a abordagem BERT-Large supera BERT-Base. Além disso, a arquitetura LSTM não oferece nenhum ganho, no entanto, a combinação de

CRF traz bom desempenho. Além disso, os modelos monolíngues superam os modelos multilíngues para NER em português. Assim, os melhores resultados foram obtidos com um modelo monolíngue treinado especificamente para o português. Embora a abordagem monolíngue tenha um desempenho melhor, o custo computacional de treinar o modelo em português é muito maior do que usar um modelo multilíngue pré-treinado.

Embora os autores forneçam uma abordagem SOTA monolíngue para NER em português, uma questão permanece: é possível que os modelos NER multilíngues possam superar os modelos unilíngues, particularmente para idiomas complexos, como o português?

Abordagens com CRF (LAFFERTY; MCCALLUM; PEREIRA, 2001) incorporado à arquitetura tem demonstrado resultados importantes na tarefa NER, principalmente se associadas a outras soluções de classificação, como é o caso da abordagem (PIROVANI, 2019), onde CRF está associado à uma camada *Local Grammar* (LG) (GROSS, 1997). Como é possível observar na Figura 5, a etapa de classificação de *tokens* foi implementada fazendo uso de uma rede CRF.

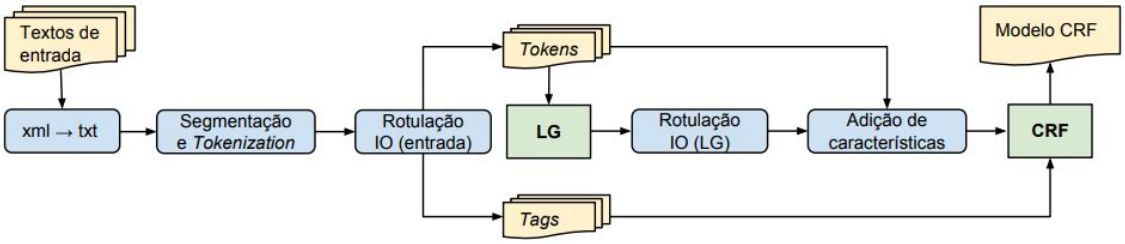

Figura 5 – Metodologia utilizada para NER associando LG e CRF.
Fonte: (PIROVANI, 2019)

O funcionamento da metodologia demonstrada na Figura 5 possui como módulos centrais a camada LG e a camada CRF. Na etapa LG realiza-se a pré-rotulação dos *tokens* de entrada, através de uma gramática local que concentra um conjunto de regras linguísticas. Ao final do *pipeline*, os dados são entregues à camada CRF da abordagem, onde ocorre o reconhecimento de entidades final, reforçando a flexibilidade de utilização de redes CRF no problema de reconhecimento de entidades, em diferentes contextos.

Conforme discutimos no capítulo 1, as abordagens e conjuntos de dados em NLP, em particular, na tarefa NER, são majoritariamente desenvolvidas com foco no idioma inglês, o que torna o desafio de desenvolver *pipelines* de extração de entidades

para o português mais restrito. Em (SILVA et al., 2020b), os autores desenvolveram uma metodologia de construção de um novo corpus em português, que poderia ser, por exemplo, uma alternativa ao HAREM (SANTOS; CARDOSO, 2007), que foi utilizado por este trabalho.

A metodologia proposta em (SILVA et al., 2020b) e evoluída com melhorias no corpus e na metodologia, conforme descrito em (SILVA et al., 2020a), utilizou arquitetura BiLSTM para NER, uma solução com resultados consistentes em outros trabalhos relacionados à NER, para que fosse possível treinar e executar NER com o *dataset* proposto, entretanto com o foco principal na construção do conjunto de dados através de extração e mapeamento das entidades, utilizando conteúdo de textos jornalísticos na estruturação desse corpus. Os experimentos com o corpus proposto, utilizado pela arquitetura BiLSTM escolhida, obtiveram resultados com precisão em torno de 85%.

Ainda com foco em conjuntos de dados consistentes para a execução de NER em português, o trabalho de (PIRES, 2017), fez uma análise comparativa de diferentes abordagens para a implementação de NER em português, com o objetivo de identificar as melhores abordagens, mas, além disso, fez a proposição de um *dataset* derivado do HAREM (SANTOS; CARDOSO, 2007), denominado de SIGARRA News Corpus, um conjunto de dados que possui 905 notícias e 12644 rotulações de entidades, derivado do HAREM, alcançando o melhor resultado em torno de 86% de *F-measure*.

4 METODOLOGIA

Nesta seção, apresentamos nossa abordagem de *embeddings* de texto multilíngues baseada em *transformers* para NER em português. Primeiramente, apresentamos um conjunto de dados abrangente que foi construído para melhorar o treinamento de modelos NER para o português. Em segundo lugar, apresentamos a arquitetura de nossa abordagem proposta.

4.1 HAREM

O conjunto de dados HAREM foi desenvolvido com o objetivo de realizar a avaliação de sistemas que executam a atividade de NER, através de uma avaliação conjunta, foi construído motivado para uma série de iniciativas com o objetivo de medir a eficácia de sistemas que reconhecem entidades em português. (SANTOS; CARDOSO, 2007)

Avaliação conjunta (SANTOS, 2007), no contexto do HAREM, foi uma adaptação do modelo de trabalho chamado *evaluation contest* (SANTOS et al., 2006), para que diferentes iniciativas no desenvolvimento de sistemas NER pudessem, de certa forma, competir entre si, utilizando o mesmo conjunto de dados, mas também que o conjunto de dados pudesse ser construído de forma colaborativa e evolutiva, como ocorreu ao longo das diferentes versões do HAREM.

Conforme apresentado na Figura 6, que representa a segunda versão do *dataset*, que apesar das particularidades de cada versão, a estrutura dos dados é organizada em categorias principais, denominadas entidades, localizadas ao centro da imagem. Além disso, os termos do conjunto pode ser divididos em tipos, que por sua vez podem ser segmentados, ainda, em subtipos, conforme descrito em (FREITAS et al., 2010).

Considerando a divisão de tipos e subtipos reapresentadas na Figura 6, é importante fazer a ressalva que para um modelo multilíngue que demanda um conjunto de dados robusto, utilização dos tipos e subtipos realizaria uma distribuição muito pulverizada dos dados, o que provavelmente causaria impacto no treinamento e consequentemente nos resultados. Desta maneira, mantivemos a utilização das categorias, que concentram um maior número de expressões rotuladas por categoria.

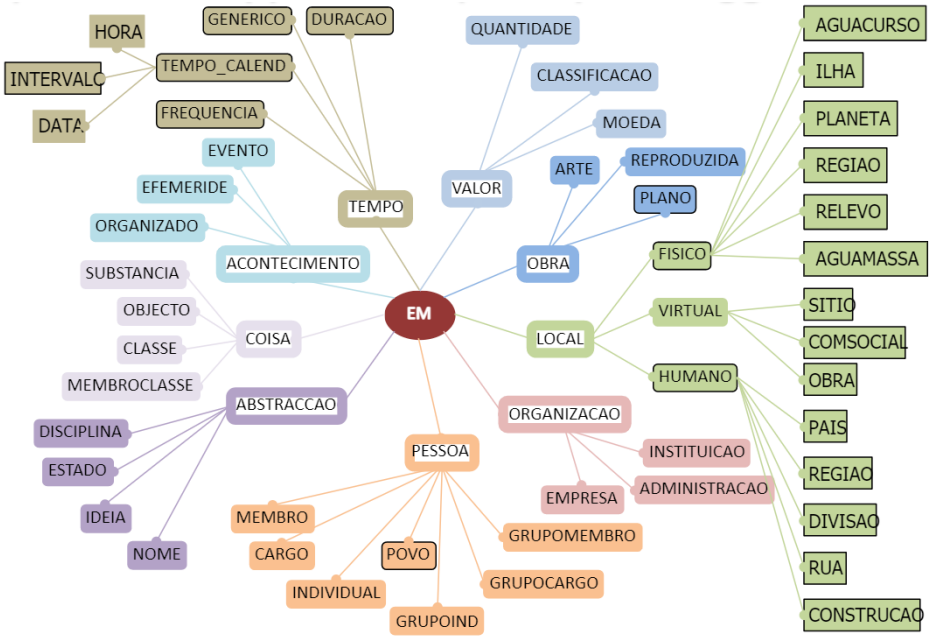

Figura 6 – Estrutura de Entidades - HAREM.
Fonte: (FREITAS et al., 2010)

4.2 Conjunto de Dados de Treinamento

Para melhorar o treinamento de modelos NER multilíngues para o português, construímos um conjunto de dados abrangente do HAREM (SANTOS; CARDOSO, 2007). HAREM[1] é um conjunto de dados anotado manualmente utilizado para avaliar o desempenho dos sistemas de informação para o reconhecimento de entidades nomeadas em português. HAREM é amplamente utilizado por várias abordagens de NLP relatadas na literatura (SOUZA; NOGUEIRA; LOTUFO, 2019; CASTRO; SILVA; SOARES, 2018; OLIVEIRA; CARDOSO, 2009; FERNANDES; CARDOSO; OLIVEIRA, 2018; CONSOLI; VIEIRA, 2019; PIRES, 2017). Em particular, o conjunto de dados HAREM tem as seguintes divisões:

- "CD Primeiro HAREM": 129 documentos e 80.060 palavras.
- "CD Segundo HAREM": 129 documentos e 147.991 palavras.
- "Mini-HAREM CD": 128 documentos e 54.074 palavras.

Todas as divisões HAREM (SANTOS; CARDOSO, 2007) foram unidas em um único conjunto de dados de treinamento unificado. Originalmente, algumas expressões

[1]Disponível em http://www.linguateca.pt

em HAREM são ambíguas, ou seja, algumas delas têm dois rótulos de entidade com significados diferentes. A Tabela 6 apresenta as categorias e definições de entidades que eventualmente possa pertencer a cada grupo.

Categoria	Descrição
O	Categoria para palavras em geral, que não são entidades
ABSTRAÇÃO	Imaginação, exemplos fictícios, projeções
OUTRO	Entidades diversas, que estão relacionadas por tipos
LOCAL	Cidades, ruas, países, distritos
ACONTECIMENTO	Eventos, fatos históricos, episódios relevantes
TEMPO	Datas, intervalos de tempo, épocas, meses ou anos
PESSOA	Nomes de Pessoas
OBRA	Livros, esculturas
ORGANIZAÇÃO	Empresas e instituições
VALOR	Valores financeiros, moeda
OBJETO	Diferentes tipos de objetos

Tabela 6 – Tipos de Entidades existentes no *dataset* HAREM

Para construir o conjunto de dados de treinamento unificado, escolhemos a primeira classificação descrita no conjunto de dados HAREM, descartando a segunda. Assim, todas as expressões foram classificadas em um único rótulo de entidade. Além disso, a estrutura do parágrafo foi convertida em sentenças menores para que o algoritmo baseado em BERT possa receber os dados de entrada em um formato apropriado. Parágrafos de até 256 *tokens* foram automaticamente convertidos em frases, além disso, os parágrafos foram divididos com rótulos de entidade, que também foram incorporadas ao conjunto de dados de treinamento unificado por se tratar de um treinamento supervisionado.

Os parágrafos grandes, com quantidade de *tokens* acima de 256, foram quebrados em frases de tamanho 256, a partir dos finalizadores de frases, por exemplo, os símbolos "!"e ".", pela característica gramatical que indica fim de uma frase com estrutura completa ao ser finalizada por esses símbolos.

4.3 Arquitetura Proposta

A abordagem multilíngue proposta para NER em português pode usar vários *embeddings* de texto baseados em transformers. Em particular, implementamos e avaliamos BERT (DEVLIN et al., 2018), RoBERTa (Liu et al., 2019) e DistilBERT (SANH et al., 2019).

A Figura 7 apresenta a arquitetura da abordagem proposta. Em alto nível, existem quatro etapas de processamento:

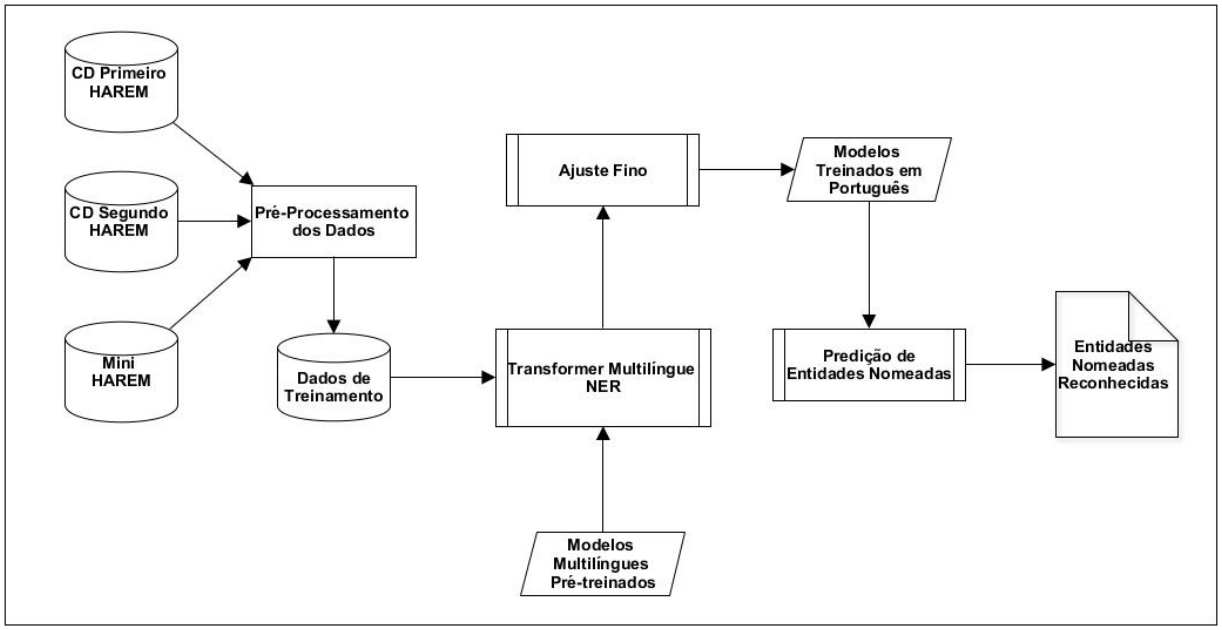

Figura 7 – A arquitetura da abordagem de *embeddings* de texto multilíngue baseada em transformer, proposta para NER em português. Fonte: O Autor

- Pré-processamento do conjunto de dados;
- Transformer NER multilingue;
- Ajuste Fino
- Predição de Entidades Nomeadas

Na etapa de pré-processamento do conjunto de dados, nossa abordagem constrói o conjunto de dados de treinamento conforme descrito na seção 4.2, originalmente no formato *Extensible Markup Language* (xml), removendo as ambiguidades originais no HAREM, padronizando os dados em sentenças dentro do padrão BERT, que recebe como entrada arquivos no formato *Tab-separated values* (tsv) e consolidação em um único arquivo de dados. Essa conversão obtém os arquivos originais, elimina frases duplicadas, e realiza a listagem dos termos, um a cada linha, indicando na mesma linha qual é a sua classificação, conforme apresentado na Figura 8. A linha é composta pelo termo e sua classificação, separados por tabulação. No exemplo apresentado na Figura 8, a entidade "Alemanha"está classificada como um local. Quando a entidade possui mais de um termo, como é o caso de "John Gutemberg", a primeira palavra tem o prefixo "B"na sua classificação e as demais palavras o prefixo "I".

```
A O
imprensa O
, O
inventada O
na O
Alemanha B–LOCAL
por O
John B–PESSOA
Gutenberg I–PESSOA
, O
foi O
importante O
na O
divulgacao O
d O
estas O
ideias O
. O
```

Figura 8 – Exemplo de parágrafo convertido para o conjunto de treinamento. Fonte: O Autor

Na segunda etapa, nossa abordagem seleciona o modelo multilíngue baseado em *transformers* para NER, instanciando-os no mecanismo de processamento e carregando os modelos multilíngues pré-treinados para a geração dos *embeddings* de texto.

Na etapa de ajuste fino, nossa abordagem define os hiper parâmetros do modelo multilíngue para a tarefa NER, gerando o modelo NER final, treinado utilizando as sentenças em português e tornando-o, portanto, especializado em predições do idioma em português.

Finalmente, na etapa de predição, nossa abordagem carrega o modelo ajustado em português, recebe todas as sentenças a serem avaliadas e gera uma saída final com as entidades nomeadas reconhecidas a partir das sentenças de entrada.

O pipeline funciona de forma flexível para que, caso uma nova versão do conjunto de dados HAREM seja publicada, seja possível incorporá-la ao conjunto de dados de treinamento, preservando o conteúdo original e ampliando o volume de dados disponíveis para modelos de treinamento e teste. Da mesma forma, embora três abordagens de *Transformers* tenham sido usadas inicialmente em nossos experimentos, também

é possível conectar novas abordagens baseadas em *Transformers* sem nenhum impacto no fluxo de trabalho de processamento.

5 EXPERIMENTOS

Nesta seção, apresentaremos os experimentos que realizamos para avaliar a abordagem proposta, incluindo a parametrização da abordagem, procedimentos e resultados. Em particular, os experimentos respondem às seguintes questões de pesquisa:

1. Qual a eficácia de cada um dos algoritmos multilíngues baseados em BERT para NER em português?
2. Qual é o desempenho de nossa abordagem multilíngue em comparação com a abordagem monolíngue SOTA para NER em português?

5.1 Treinamento e Teste

Para analisar a influência do tamanho do conjunto de treinamento nos resultados, faseamos os experimentos em cenários de treinamento distintos:

- 70% dos dados para treinamento e 30% dos dados para teste;
- 80% de dados para treinamento e 20% de dados para teste;
- 90% de dados para treinamento e 10% de dados para teste;
- 95% de dados para treinamento e 5% de dados para teste.

Em cada um dos cenários, avaliamos BERT (DEVLIN et al., 2018), XLM-RoBERTa (LAMPLE; CONNEAU, 2019) e DistilBERT (SANH et al., 2019), também realizando um ajuste fino para a tarefa NER . Para uma comparação coerente, o mesmo conjunto de dados de treinamento e parâmetros de configuração foram usados para cada algoritmo, todos baseados em BERT.

Devido à quantidade limitada de dados, *batches* maiores foram descartados, por exemplo, *batches* com 512 *tokens*. *Batches* com número elevado de *tokens* implicam na redução do número de sentenças enviadas para a entrada do algoritmo, impactando negativamente no desempenho. Assim, os *batches* de 128 e 256 tornaram-se mais adequados para nossos experimentos. *Batches* menores que 128 podem causar problemas de truncamento, ou seja, as frases são recortadas em partes menores, separando-as, gerando mais perda de dados ou frases com estruturação linguística inadequada.

Abordagens baseadas em transformers, particularmente BERT, geralmente requerem poucas interações para convergir em um modelo capaz de fornecer resultados eficientes (WOLF et al., 2019). Testamos diferentes épocas para finalmente definir este parâmetro como 3, para um melhor equilíbrio entre o tempo de treinamento e o desempenho do modelo. A Tabela 7 apresenta o desempenho de nossa abordagem multilíngue usando diferentes algoritmos baseados em BERT com múltiplas variações de conjunto de treinamento e 3 épocas de treinamento.

Abordagem	**Treino (%)**	**SEQ_SIZE**	**Precisão (%)**	**Revocação (%)**	**F1 (%)**
BERT-Base	95	128	85.00	86.80	85.90
BERT-Base	95	256	85.70	86.30	86.00
DistilBERT	95	128	77.10	82.90	79.90
DistilBERT	95	256	78.50	83.00	80.70
XML-RoBERTa	95	128	**88.00**	87.60	**87.80**
XML-RoBERTa	95	256	86.30	**88.40**	87.30
BERT-Base	90	128	67.00	74.20	70.40
BERT-Base	90	256	68.60	75.40	71.80
DistilBERT	90	128	62.30	68.20	65.10
DistilBERT	90	256	62.60	69.30	65.80
XML-RoBERTa	90	128	73.00	78.60	75.70
XML-RoBERTa	90	256	**74.60**	**79.80**	**77.10**
BERT-Base	80	128	66.40	69.90	68.10
BERT-Base	80	256	68.30	71.20	69.70
DistilBERT	80	128	59.10	64.60	61.70
DistilBERT	80	256	60.80	64.70	62.70
XML-RoBERTa	80	128	**67.90**	70.90	69.40
XML-RoBERTa	80	256	67.90	**71.50**	**69.70**
BERT-Base	70	128	61.40	62.30	61.80
BERT-Base	70	256	62.50	64.40	63.40
DistilBERT	70	128	58.00	59.50	58.80
DistilBERT	70	256	59.30	61.10	60.20
XML-RoBERTa	70	128	**64.40**	**64.80**	**64.60**
XML-RoBERTa	70	256	64.10	64.80	64.40

Tabela 7 – Desempenho da abordagem multilíngue usando algoritmos baseados em BERT com variações no conjunto de dados e 3 épocas de treinamento. Fonte: O Autor

Conforme apresentado na Tabela 7, observamos que XML-RoBERTa supera BERT e DistilBERT em diferentes cenários. Particularmente, o volume de dados de treinamento impacta o desempenho de todos os algoritmos baseados em BERT, com XML- RoBERTa superando BERT-Base em 2,68 % na precisão, 1,84 % na revocação e 2,09 % no F1, também superando DistilBERT em 12,10 % em precisão, 6,50 % em revocação e 8,79 % em F1, considerando 95 % do cenário de treinamento. Além disso,

podemos observar que as diferenças no desempenho de XML- RoBERTa com *batches* de 128 e 256 são desprezíveis (0,57 % em F1). Relembrando nossa primeira questão de pesquisa, estes resultados atestam a eficácia de nossa abordagem multilíngue RoBERTa para NER em português.

5.2 Resultados

A tabela 8 apresenta o desempenho da abordagem de *embeddings* de texto baseada em transformer monolíngue SOTA relatada na literatura e, também, nossa proposta de abordagem de *embeddings* de texto baseada em transformer multilíngue para NER em português. E possível observar na tabela 8 que nossa abordagem multilíngue (XML-RoBERTa) supera a melhor abordagem monolíngue (PT-BERT-Large -CRF) no cenário completo, com ganhos de 9,89% em precisão, 13,31% em revocação e em 11,60% em F1. Mesmo considerando o cenário seletivo (melhor) para a abordagem monolíngue, os ganhos ainda são significativos de 3,74% na precisão, 7,19% no recall e 5,47% na F1. Relembrando nossa segunda questão de pesquisa, esses experimentos mostram que as abordagens de *embeddings* de texto multilíngues baseadas em transformers, ajustadas com um grande conjunto de dados, superam os modelos baseados em transformer SOTA treinados especificamente para o português.

Abordagem	**Precisão**	**Revocação**	**F1**
Monolígue (Cenário Completo)			
LSTM-CRF	72.78	68.03	70.33
BiLSTM-CRF+FlairBBP	74.91	74.37	74.64
ML-BERT-Base-CRF	74.82	73.49	74.15
PT-BERT-Base-CRF	78.60	76.89	77.73
PT-BERT-Large-CRF	80.08	77.31	78.67
Monolígue (Cenário Seletivo)			
LSTM-CRF	78.26	74.39	76.27
BiLSTM-CRF+FlairBBP	83.38	81.17	82.26
ML-BERT-Base-CRF	80.10	78.78	79.44
PT-BERT-Base-CRF	83.89	81.50	82.68
PT-BERT-Large-CRF	84.82	81.72	83.24
Multilíngue (Cenário Completo)			
DistilBERT	78.50	83.00	80.70
BERT-Base	85.70	86.30	85.90
XLM-RoBERTa	**88.00**	**87.60**	**87.80**

Tabela 8 – Desempenho do SOTA monolíngue e as abordagens de *embeddigns* de texto multilíngues propostas, baseadas em transformers, para NER em português

Abordagens baseadas em *transformers* multilíngues para NER tornam-se interessantes em cenários onde a quantidade de recursos computacionais é limitada para

treinar abordagens monolíngues específicas para um idioma, mas a quantidade de dados de treinamento é abundante para o ajuste fino. Além disso, a etapa de ajuste fino pode ser generalizada para qualquer abordagem multilíngue baseada em BERT. Portanto, ALBERT (LAN et al., 2019) e BART (LEWIS et al., 2019), entre outras, podem ser facilmente implementadas em nossa abordagem baseada em transformers tão logo seja disponibilizada a versão multilíngua dessas abordagens, da mesma forma que implementamos DISTILBERT (SANH et al., 2019) e ROBERTA (Liu et al., 2019).

6 CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho, avaliamos a eficácia de *embeddings* de texto multilíngues baseados em *transformers* para o reconhecimento de entidades nomeadas em português. Particularmente, aprimoramos nossa abordagem multilíngue usando um grande conjunto de dados de texto em português e realizamos experimentos comparando nossa abordagem multilíngue com a abordagem monolíngue de última geração, com modelo treinado especificamente para o português.

Os resultados mostraram que nossa abordagem de *embeddings* de texto baseada em *transformer* multilíngue superou a abordagem do estado da arte, alcançando 88,0% de precisão e 87,8% em F1 no reconhecimento de entidades nomeadas para o português, com ganhos de até 9,89% de precisão e 11,60% em F1. Além disso, mesmo considerando um cenário seletivo em que a abordagem do estado da arte teve um desempenho melhor, nossa abordagem multilíngue superou o estado da arte em 3,74% de precisão e 5,47% em F1. Assim, nossos experimentos mostraram que modelos de linguagem genérica multilíngues pré-treinados, baseados em BERT, e ajustados com um conjunto de dados maior podem superar os modelos de linguagem específicos de um único idioma que requerem muito tempo e recursos computacionais para serem treinados.

Em trabalhos futuros, pretendemos avaliar o impacto do tamanho do conjunto de dados unificado sobre a eficácia do modelo NER, bem como melhorar os algoritmos baseados em *transformers* para que seja possível ajustar os *batches* para tamanhos menores, por exemplo 64, permitindo aumentar o número de sentenças analisadas e possivelmente obter resultados ainda melhores.

Ainda com foco no tamanho e na variedade do conjunto de dados, gostaríamos de avaliar a possibilidade de comparar o desempenho da abordagem aplicando-a em outros conjuntos de dados em português, por exemplo, o dataset proposto em (SILVA et al., 2020b, 2020a). Por sua similaridade ao perfil de escrita adotado no HAREM, seria possível considerar a possibilidade de incluir na abordagem proposta uma tarefa de fusão entre essa base com o HAREM, reforçando e endereçando nossa questão de pesquisa na direção de considerar um conjunto de dados maior que aproxime os resultados de abordagens multilíngues aos de abordagens monolíngues desenvolvidas especificamente para o português.

Além disso, de forma semelhante à abordagem monolíngue de última geração, pretendemos adicionar uma camada de CRF à nossa abordagem multilíngue, o que

pode melhorar ainda mais a precisão, conforme ocorrido no SOTA, que possui características semelhantes a este trabalho, tanto na metodologia utilizada, quanto nos dados de entrada, além das abordagens em comum entre as duas propostas.

REFERÊNCIAS

ARKHIPOV, M. et al. Tuning multilingual transformers for language-specific named entity recognition. In: Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing. [S.l.: s.n.], 2019. p. 89–93.

BENGIO, Y. et al. A neural probabilistic language model. The Journal of Machine Learning Research, v. 3, p. 1137–1155, 2003.

BOJANOWSKI, P. et al. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, v. 5, p. 135–146, 2017.

BORTHWICK, A. E. A maximum entropy approach to named entity recognition. 1999. Tese (Doutorado) — New York University, USA.

CASTRO, P. V. Q. de; SILVA, N. F. F. da; SOARES, A. da S. Portuguese named entity recognition using LSTM-CRF. In: Proceedings of the International Conference on Computational Processing of the Portuguese Language. [S.l.: s.n.], 2018. p. 83–92.

CONSOLI, B.; VIEIRA, R. Multidomain contextual embeddings for named entity recognition. In: Proceedings of the Iberian Languages Evaluation Forum. [S.l.: s.n.], 2019. p. 434–441.

DEVLIN, J. et al. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805, 2018.

FERNANDES, I.; CARDOSO, H. L.; OLIVEIRA, E. Applying deep neural networks to named entity recognition in portuguese texts. In: Proceedings of the 5th International Conference on Social Networks Analysis, Management and Security. [S.l.: s.n.], 2018. p. 284–289.

FREITAS, C. et al. Second HAREM: Advancing the state of the art of named entity recognition in portuguese. In: Proceedings of the 7th International Conference on Language Resources and Evaluation. [S.l.: s.n.], 2010.

GOLDBERG, Y. A primer on neural network models for natural language processing. Journal of Artificial Intelligence Research, v. 57, n. 1, p. 345–420, 2016.

Gomez, A. N. et al. The Reversible Residual Network: Backpropagation Without Storing Activations. CoRR, 2017.

GROSS, M. The construction of local grammars. Finite-state language processing, 1997.

GUTMANN, M. U.; HYVÄRINEN, A. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. The Journal of Machine Learning Research, v. 13, n. 1, p. 307–361, 2012.

HINTON, G.; VINYALS, O.; DEAN, J. Distilling the knowledge in a neural network. In: Proceedings of the NIPS Deep Learning and Representation Learning Workshop. [S.l.: s.n.], 2015.

KIROS, R. et al. Skip-Thought vectors. In: Proceedings of the 29th Conference on Neural Information Processing Systems. [S.l.: s.n.], 2015. p. 1532–1543.

LAFFERTY, J. D.; MCCALLUM, A.; PEREIRA, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning. [S.l.: s.n.], 2001. p. 282–289. ISBN 1558607781.

LAMPLE, G.; CONNEAU, A. Cross-lingual language model pretraining. CoRR, abs/1901.07291, 2019.

LAN, Z. et al. ALBERT: A lite BERT for self-supervised learning of language representations. CoRR, 2019.

LEVY, O.; GOLDBERG, Y. Dependency-based word embeddings. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.], 2014. p. 302–308.

LEWIS, M. et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. CoRR, abs/1910.13461, 2019.

Liu, Y. et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR, 2019.

MARSH, E.; EPERZANOWSKI, D. MUC-7 evaluation of IE technology: Overview of results. In: Proceedings of the 7th Message Understanding Conference. [S.l.: s.n.], 1998.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. CoRR, abs/1301.3781, 2013.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. [S.l.: s.n.], 2013. p. 3111–3119.

MORIN, F.; BENGIO, Y. Hierarchical probabilistic neural network language model. In: Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics. [S.l.: s.n.], 2005. p. 246–252.

OLIVEIRA, H. G.; CARDOSO, N. Sahara: An online service for harem named entity recognition evaluation. In: Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology. [S.l.: s.n.], 2009. p. 171–174.

PENNINGTON, J.; SOCHER, R.; MANNING, C. GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. [S.l.: s.n.], 2014. p. 1532–1543.

PETERS, M. E. et al. Deep contextualized word representations. CoRR, abs/1802.05365, 2018.

PIRES, A. R. O. Named entity extraction from Portuguese web text. 2017. Dissertação (Mestrado) — Porto University, Portugal.

PIROVANI, J. P. C. CRF+LG: uma abordagem híbrida para o reconhecimento de entidades nomeadas em português. 2019. Tese (Doutorado) — UNESP, Brazil.

SANH, V. et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR, abs/1910.01108, 2019.

SANTOS, D. Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa. [S.l.]: Instituto Superior Técnico Press, 2007.

SANTOS, D.; CARDOSO, N. (Ed.). Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. [S.l.]: Linguateca, 2007.

SANTOS, D. et al. HAREM: An advanced NER evaluation contest for Portuguese. In: Proceedings of the 5th International Conference on Language Resources and Evaluation. [S.l.: s.n.], 2006.

SHAZEER, N. et al. Mesh-TensorFlow: Deep learning for supercomputers. In: BENGIO, S. et al. (Ed.). Advances in Neural Information Processing Systems. [S.l.]: Curran Associates, Inc., 2018. v. 31.

SILVA, R. de A. et al. An improved ner methodology to the portuguese language. Mobile Networks and Applications, p. 1–7, 2020.

SILVA, R. de A. et al. A new entity extraction model based on journalistic brazilian portuguese language to enhance named entity recognition. In: MUGNAINI, R. (Ed.). Data and Information in Online Environments. Cham: Springer International Publishing, 2020. p. 53–63.

SOUZA, F.; NOGUEIRA, R. F.; LOTUFO, R. de A. Portuguese named entity recognition using BERT-CRF. CoRR, abs/1909.10649, 2019.

VASWANI, A. et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. [S.l.: s.n.], 2017. p. 6000–6010.

WOLF, T. et al. HuggingFace's Transformers: State-of-the-art natural language processing. CoRR, 2019.

XUN, G. et al. A correlated topic model using word embeddings. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. [S.l.: s.n.], 2017. p. 4207–4213.

YANG, Z. et al. XLNet: Generalized autoregressive pretraining for language understanding. CoRR, abs/1906.08237, 2019.