DOI: 10.5748/19CONTECSI/PSE/DSC/7035

WORD EMBEDDING FOR UNKNOWN WORDS: ADDING NEW WORDS INTO BERT'S VOCABULARY

EMBEDDING PARA PALAVRAS DESCONHECIDAS: ADICIONANDO PALAVRAS NOVAS AO VOCABULÁRIO DO MODELO BERT

Letícia Silveira Artese ; https://orcid.org/0000-0003-3380-3494
Universidade Federal de Santa Catarina - UFSC

Daniel Maciel ; https://orcid.org/0000-0002-1706-6423
Universidade Federal de Santa Catarina - UFSC

Alexandre Leopoldo Gonçalves ; https://orcid.org/0000-0002-6583-2807
Universidade Federal de Santa Catarina - UFSC

WORD EMBEDDING FOR UNKNOWN WORDS: ADDING NEW WORDS INTO BERT'S VOCABULARY

EMBEDDING PARA PALAVRAS DESCONHECIDAS: ADICIONANDO PALAVRAS NOVAS AO VOCABULÁRIO DO MODELO BERT

ABSTRACT
In natural language processing, dealing with the dynamics of languages, such as the arisen of new words, can be a challenge to models. In deep learning models, when a word is not presented in the training dataset, it is not known by the model and, therefore, considered out of vocabulary (OOV). Although many models manage to get around this barrier, sometimes it is necessary to learn the embedding of a new word. In this sense, a method is presented to obtain a dynamic contextual vector representation of a new word based in the BERT language model. To evaluate the method, we took the case of the arisen of the word 'voip' in scientific publications, obtaining an embedding close to 'telecommunications' and 'signalling', some of the main words with significance in relation to the context of the word of study, demonstrating that the proposed method offers an efficient way to obtain embeddings for new words.

Keywords: NLP, word embedding, new word, rare word, OOV

RESUMO
No processamento de linguagem natural, lidar com a dinamicidade das línguas, como o surgimento de novas palavras, pode ser um desafio aos modelos. Visto que, em modelos de aprendizado profundo quando uma palavra não é apresentada na etapa de treinamento ela não é conhecida pelo modelo e, portanto, considerada fora do vocabulário (OOV). Apesar de muitos modelos conseguirem contornar essa barreira, às vezes se faz necessário aprender o *embedding* de novas palavras. Neste sentido, apresenta-se um método para obtenção da representação vetorial contextual dinâmica de palavras novas a partir do modelo de linguagem BERT. Na avaliação do método, foi utilizado o caso do surgimento da palavra '*voip'* em artigos científicos, obtendo um *embedding* próximo de *'telecommunications'* e *'signalling'*, algumas das principais palavras com significância em relação ao contexto da palavra de estudo, demonstrando que o método proposto oferece uma maneira eficiente para obter *embeddings* para palavras novas.

Palavras-chave: PLN, embedding de palavra, palavra nova, palavra rara, OOV

1. INTRODUÇÃO

As abordagens existentes para aprender a representação de palavras geralmente assumem que há ocorrências suficientes para cada palavra no corpus, de modo que a representação das palavras pode ser estimada com precisão a partir de seus contextos. No entanto, em cenários do mundo real, palavras fora do vocabulário (*Out Of Vocabulary* - OOV) que não aparecem no corpus de treinamento surgem com frequência (Hu, Chen, Chang, & Sun, 2019). A língua é viva e muda rapidamente. Apresenta um curso que as transforma constantemente em alternativas mais aderentes aos grupos e sociedades das quais fazem parte (Bichakjian, 2017). Constitui-se como elemento em constante construção, adaptação e aperfeiçoamento (Gontier, 2017). Enquanto algumas palavras deixam de ser usadas, outras são criadas a partir de novas formas de uso ou modificações das existentes (Hombaiah, Chen, Zhang, Bendersky, & Najork, 2021). Essas palavras desconhecidas podem estar relacionadas, por exemplo, a termos técnicos de domínios de conhecimento específico ou, também, textos informais de internet, onde ocorrem gírias, neologismos e erros de digitação que podem ser consideradas palavras desconhecidas. Uma representação semântica adequada é requisito essencial para tarefas de Processamento de Linguagem Natural (*Natural Language Processing* - NLP). Portanto, encontrar boas representações para palavras OOV representa um desafio relevante.

Modelos de linguagem pré-treinados em extensos corpus de texto apresentam bons resultados em análises e tarefas relacionadas ao processamento contextualizado de palavras e sentenças (Devlin, Chang, Lee, & Toutanova, 2019; Tai, Kung, & Dong, 2020). Modelos de linguagem que apresentam arquitetura com base em '*transformer models' e 'attention mechanisms'* (Vaswani et al., 2017) como BERT® (*Bidirectional Encoder Representations from Transformers*) (Devlin et al., 2019) e GPT-3® (Brown et al., 2020), representam o estado da arte em NLP e vem sendo usados por apresentarem resultados superiores às técnicas precedentes de modelos sequenciais, como, LSTM (*Long Short-T erm Memory*), RNN (*Recurrent Neural Network*), GRU (*Gated Recurrent Unit*), entre outros. Tendo seu sucesso atrelado ao conceito de transferência de conhecimento (*transfer learning*) permitindo que modelos treinados em grandes corpus de textos genéricos possam ser aplicados em tarefas específicas sem a necessidade de replicação extensa do treinamento (Qiu et al., 2020; Subakti, Murfi, & Hariadi, 2022).

Além disso, esses modelos de linguagem pré-treinados superam modelos NLP antes largamente usados, como o Word2Vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), por oferecerem *embeddings* contextualizados dinâmicos. Abordagens como Word2vec (*C-BOW* ou *Skip-gram)* fornecem *embeddings* estáticos, pois exibem um único vetor no espaço semântico. Por outro lado, abordagens como o BERT fornecem *embeddings* com semântica dinâmica, contextualizada, isto é, são capazes de gerar *embeddings* diferentes para uma mesma palavra que apresente contextos distintos. Um progresso significativo para o processamento de linguagem natural considerando que palavras frequentes, de uso corrente, tendem a apresentar mais sentidos, de acordo com o Princípio da Versatilidade Econômica das Palavras (Zipf, 1949).

O BERT é um modelo de linguagem pré-treinado em um grande corpus de texto (Wikipedia® e BooksCorpus®) executando duas tarefas: linguagem mascarada (*Masked Language Modelling - MLM*) e previsão da próxima sentença (*Next Sentence Prediction NSP)* para que o *embedding* obtido pelo modelo BERT ocorra de acordo com o contexto da frase na qual ela se encontra. Além disso, o modelo aprende as representações em um nível de 'subpalavra', devido à implementação do algoritmo de tokenização, *sub-word tokenizer, WordPieces* (Wu et al., 2016), possibilitando que o vocabulário base do modelo BERT conste com mais de trinta mil *tokens* (Devlin et al., 2019). Essa flexibilização para a

representação dos textos em *tokens*, que representam tanto palavras quando sub-palavras, possibilita que o modelo BERT possa lidar com palavras fora do vocabulário (OOV).

Ademais, a tokenização em subpalavras baseia-se no princípio de que palavras usadas com frequência não devem ser segmentadas em subpalavras, e palavras menos usuais devem ser decompostas em subpalavras. Dessa forma, dependendo da aplicabilidade, quando o problema apresenta grande quantidade de palavras desconhecias, consideradas OOV, como palavras de domínio específico; palavras novas, como novos termos de nomenclatura científica ou neologismos; e palavras raras, aquelas que apresentam baixa frequência de ocorrência no conjunto de dados. Esses problemas muitas vezes não apresentam desempenho satisfatório, pois seus *embeddings* tem pouca representatividade semântica (Hu et al., 2019).

Para atender o problema de palavras OOV, com a finalidade de lidar com palavras de domínio específico, desenvolveram-se modelos BERT dedicados, gerados a partir do treinamento do zero do modelo BERT padrão em extensos corpus dedicados a temas particulares, como área jurídica ou médica. Resultando em uma gama de modelos como SciBERT (Beltagy, Lo, & Cohan., 2019), BioBERT (Lee et al., 2019), PubMedBERT (Gu et al., 2021) entre outros. Todavia, treinar o modelo BERT do zero é muito caro (Strubell, Ganesh, & McCallum, 2019; Tai et al., 2020) e não descarta a possibilidade do modelo lidar com *embedding* de palavras novas ou raras. Nesse sentido, outras estratégias para contornar esse problema foram sugeridas, como a substituição de palavras OOV por sinônimos, como no algoritmo PatchBERT (Moon & Okazaki, 2020) ou a aproximação de palavras raras por radicais e ajustadas pelo contexto, como sugerido pelo BERTRAM (Schick & Schütze, 2020), porém quando o interesse está em aprender o *embedding* da palavra nova existem poucas abordagens disponíveis.

Nesse sentido, ainda é incipiente a disponibilidade de modelos que busquem satisfazer a necessidade de aprender o *embedding* de uma palavra desconhecida, nova ou rara. A linguagem apresenta rápida e constante evolução, percebida, por exemplo, na frequente adição de novas palavras aos dicionários, como foi o caso de “covid” e “zoom” em 2020 (Hombaiah et al., 2021). Como efeito, muitas vezes existe uma necessidade em obter o *embedding* desta palavra como um único *token*, especialmente, se existe um interesse em análises temporais da relação do *embedding* dessa palavra nova com o contexto. Uma vez que, além do surgimento de novos termos, palavras e expressões já consolidadas passam frequentemente por modificações semânticas à medida que seus entendimentos são aplicados em diversos contextos (*semantic shift*) (Kutuzov, Øvrelid, Szymanski, & Velldal, 2018; Hombaiah et al., 2021).

Uma das poucas abordagens que contribuem para esse problema é o modelo exBERT proposto por Tai et al. (2020). A proposta do exBERT altera a estrutura original do BERT anexando um módulo de extensão para incorporar o vocabulário do domínio específico ao mesmo tempo que faz uso dos pesos do modelo pré-treinado. No entanto, a estratégia não apresenta uma implementação clara e se mostra dependente de um grande aparato de processamento computacional e de um corpus de treinamento robusto. Nesse sentido, o método proposto neste trabalho se aproxima do trabalho de Tai et al. (2020) no intuito de buscar uma alternativa para ensinar o modelo BERT uma palavra nova, como um único *token*, e obter seu *embedding* contextual. Todavia, lidamos com a limitação de recursos computacionais e conjunto de dados reduzido.

Encerra-se a sessão de introdução dando sequência a sessão de metodologia apresentando o modelo conceitual desenvolvido. Na sessão seguinte, resultados e discussão, expõem uma instanciação do modelo. Finalizando o artigo com a sessão de conclusão.

2. METODOLOGIA

Diretrizes da metodologia *Design Science Research* (DSR) (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007) foram adotadas como base para o desenvolvimento do método proposto como solução para o problema de *embedding* de palavras novas passível de análise temporal. O modelo consiste de 3 principais etapas; (1) inclusão da palavra nova no vocabulário do modelo BERT; (2) obtenção do *embedding* a partir do corpus de treinamento pré-estabelecido; (3) extração da similaridade da palavra nova para análise temporal das variações de contexto semântico.

2.1. Inclusão da palavra nova no vocabulário BERT

A inclusão da palavra nova no vocabulário do modelo BERT pode ocorrer de duas formas, dependendo da finalidade do problema para o qual o modelo BERT será utilizado. A primeira permite a inserção direta da palavra no arquivo de vocabulário presente no diretório raiz do BERT. Este documento, o *vocab.txt*, possui todos os *tokens* de palavras conhecidas pelo modelo, incluindo subpalavras, numerais, símbolos e caracteres especiais. O arquivo possui ainda aproximadamente mil espaços não utilizados, nomeados como *'[unusedn]'*, que são reservados para a inclusão de novos *tokens*, no caso, bastando substituir essas lacunas pelas palavras nova a serem adicionadas. Essa alternativa é útil para poucas palavras alvo de interesse.

A segunda estratégia, é por meio de uma função nativa do *tokenizer*, o método *add_tokens*, que recebe como parâmetros os termos a serem adicionados ao vocabulário. Este método, todavia, não realiza alterações no arquivo do vocabulário original, *vocab.txt*, mas cria um arquivo anexo de configuração do tipo JSON, que armazena os novos *tokens* e seus respectivos *ids*. O método *add_tokens*, apesar de ser nativo, faz uma alteração no número de linhas do vocabulário, exigindo um rebalanceamento no tamanho da matriz *token-embeddings* do modelo, sendo, portanto, uma estratégia recomendada para casos que necessitam a inclusão de um grande volume de palavras.

Para execução deste trabalho, optou-se pela inclusão direta da palavra ao arquivo de vocabulário. Dessa forma, o modelo BERT passa então a reconhecer a nova palavra nova como um único *token*.

2.2. Obtenção do *embedding* a partir do corpus de treinamento pré-estabelecido

O modelo BERT pode ser implementado com duas finalidades: (i) extração de atributos (*feature-based approach*), isto é, para a obtenção de *embeddings* contextuais dinâmicos a partir de um conjunto de textos específicos, executando as tarefas padrões do modelo BERT de MLM e NSP reajustando os pesos dos vetores dos *embeddings* do dicionário do BERT para determinado contexto e reutilizá-los em outra tarefa segundo a necessidade do estudo; (ii) ajuste fino (*fine-tunning*), o qual se trata da adição de uma última camada no modelo BERT dedicada a realizar alguma tarefa em específica como classificação, identificação de entidades e outras. Aqui trabalhamos com a primeira possibilidade.

Cada um dos *tokens* do vocabulário possui um vetor de representação atribuído de acordo com o pré-treinamento do modelo. Na implementação do modelo BERT com a finalidade de extração de atributos, as 768 posições deste vetor de representação são modificadas durante a execução das tarefas padrões do modelo BERT (MLM e NSP) a

partir de um *dataset* de treinamento pré-estabelecido providenciando contextualização para a palavra nova de interesse. Dessa forma, ao final de cada *dataset* de treinamento, por exemplo, a cada conjunto de dados referentes a um ano específico, a palavra nova inserida no vocabulário na Etapa 1 possui um novo *embedding* ajustado em relação ao *dataset* anterior. O processo é repetido para quantos *datasets* forem necessários até se obter um *embedding* semanticamente representativo para a palavra nova.

2.3. Extração da similaridade da palavra nova para análise temporal das variações de contexto semântico

Tendo em vista a necessidade de constatar a coerência dos *embeddings* obtidos nos processos de extração de atributos do modelo BERT na Etapa 2, tomou-se como parâmetro de avaliação a aproximação do significado da palavra nova de interesse com demais palavras do vocabulário BERT analisando se compõem um contexto semântico coerente.

Optou-se pelo cálculo de similaridade por cossenos para verificar a proximidade das palavras, sendo possível observar após cada ajuste quais são as palavras que possuem *embeddings* mais próximos e, consequentemente, representam palavras com de um contexto significativo. Esse processo é repetido após a extração de atributos de cada *dataset*, isto é, de cada ano, elencando as *n* palavras com maior proximidade da palavra nova, e também indicando as variações dos seus significados e similaridades à medida que a quantidade de ocorrências aumenta e o aprendizado do modelo é acumulado. A Figura 1 traz uma síntese do modelo proposto.

Figura 1: síntese das etapas do modelo proposto para *embedding* de palavras novas

Fonte: autores.

3. RESULTADO

Nesta seção detalha-se o método proposto apresentando os componentes tecnológicos utilizados no desenvolvimento e como estes se interconectam nas etapas de modo a efetivamente incluir a palavra no modelo e produzir seus *embeddings* contextualizados.

Como estudo de caso para demonstrar a viabilidade do método proposto, foi utilizanda a palavra *'voip'* como uma palavra nova alvo de interesse. A tecnologia VoIP, do inglês *Voice **o**ver Internet Protocol* (em português traduzido como, V**o**z sobre IP), permite realizar chamadas de voz usando uma conexão de *Internet* de banda larga em vez de uma linha telefônica normal (ou analógica). Por se tratar de uma tecnologia criada na década de noventa, o que garante que não existam ocorrências antes deste período, e também por ser uma tecnologia que conseguiu se estabelecer e se difundir de maneira consistente, tornando-se tema frequente em publicações e conferências. Ademais, a palavra tornou-se apta para o estudo uma vez que não está nativamente inclusa no vocabulário BERT.

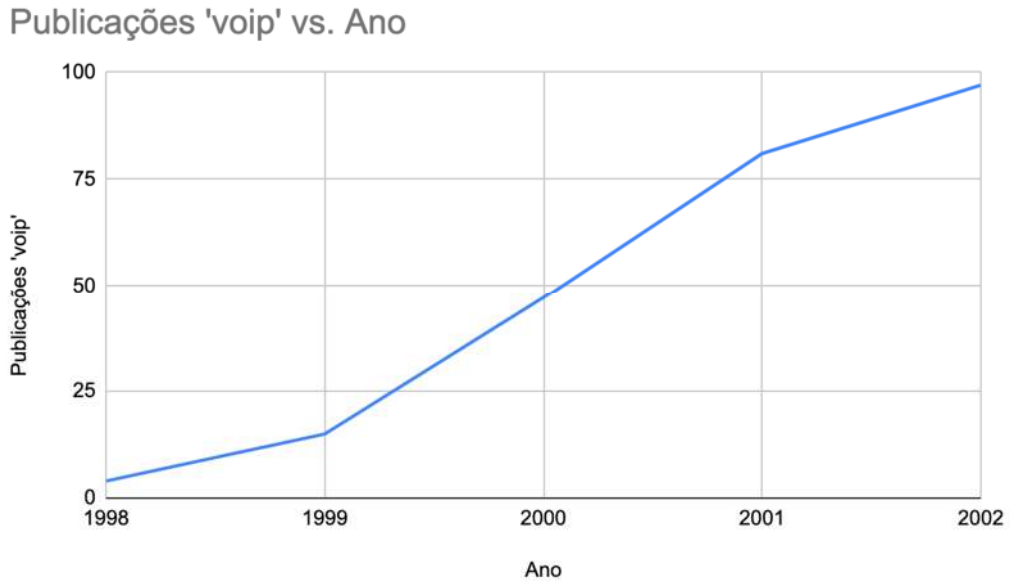
3.1. Inclusão da palavra nova no vocabulário BERT

A implementação toma como início a versão do modelo BERT *base uncased.* Utilizamos para o desenvolvimento deste estudo de caso o ambiente em nuvem Colab da Google®. Este recurso, oferecido no modelo plataforma como serviço (do inglês *Platform as a Service - PaaS*), disponibiliza simultaneamente o ambiente e a infraestrutura necessária para o desenvolvimento de *notebooks* Python®. Outrossim, oferece também máquinas virtuais com arquiteturas CPU, GPU e TPU adaptadas para o desenvolvimento de modelos de *Deep Learning*. Consoante, tendo sido estabelecida a infraestrutura necessária, a palavra *'voip'* foi incluída no vocabulário do modelo BERT utilizando o espaço não utilizado *'[unused992]'*, presente no arquivo *vocab.txt.*

3.2. Obtenção do *embedding* a partir do corpus de treinamento pré-estabelecido

Nesta etapa, antes da obtenção dos *embeddings* com o modelo BERT, foram criados os *datasets* de treinamento extraindo resumos de artigos acadêmicos e seu devido pré-processamento. Para o conjunto de dados, foram recuperados artigos na base *Scopus*, restringindo às publicações em conferências, com a palavra *'voip*' nos campos '*title*', '*abstract*' e '*keywords*'. A primeira aparição data de 1998 e encerramos o *dataset* com o ano de 2002, quando o volume de publicações começa a crescer. O Gráfico 1 apresenta o crescimento de publicações para cada ano de busca.

Gráfico 1: Publicações em conferências a partir da palavra '*voip*' na base Scopus®



Fonte: autores.

Visto que o modelo BERT necessita que o texto seja devidamente preparado, isso inclui funções de limpeza e transformação, como remoção de caracteres especiais, numerais e mudança das palavras para caixa baixa. Também, um pré-processamento nas sentenças do *dataset* para deixar todas as frases na configuração correta aceitas pela função de treinamento do BERT, como a inserção dos *tokens* especiais '[CLS]' e '[SEP]' que funcionam como indicadores do começo e do fim da sentença para o modelo. Os textos

foram divididos em sentenças, de acordo com a pontuação final de cada frase, uma vez que o modelo foi configurado para trabalhar com *strings* de até 256 caracteres. Por fim, o *dataset* de cada ano foi constituído com o número de frases, conforme a Tabela 1.

Tabela 1: Relação da quantidade de sentenças por *dataset* e ocorrências da palavra *'voip'*

Dataset	Quantidade de frases	Ocorrências da palavra *'voip'*
1998	24	7
1999	111	23
2000	279	79
2001	568	161
2002	704	195

Fonte: autores.

À medida que o número de publicações aumenta com o passar do tempo, aumenta o número de sentenças e, consequentemente, a quantidade de ocorrências da palavra alvo, neste caso *'voip'*. O *embedding* do *token* é ajustado no modelo sempre que há uma ocorrência da palavra correspondente no *dataset*, de modo que quanto maior for o número de ocorrências da palavra, mais consolidada é sua representação vetorial semântica contextual.

Após os ajustes ao nível de texto, o *dataset* é dividido em um número de épocas, definido conforme a quantidade de iterações que serão realizadas durante a execução da tarefa de extração de atributos do BERT. Para este estudo de caso, o número de épocas foi definido empiricamente, verificando quantas iterações eram necessárias para ajustar o *embedding* da palavra nova, de modo que seu significado se aproxima de palavras da mesma área, por exemplo *'voip'* e *'telecommunication'*. Verificou-se assim que 10 épocas apresentavam um resultado satisfatório. Findada esta etapa de configurações, o modelo está pronto para o treinamento.

A primeira iteração inicia-se utilizando o modelo BERT original com a palavra *'voip'* inserida e o *dataset* de 1998 utilizado como corpus de treinamento. Este processo gerou então um modelo ajustado conforme o contexto e as ocorrências das palavras nos artigos de 1998. Este novo modelo foi então usado como partida para o *dataset* de 1999, incorporando assim o aprendizado das ocorrências das palavras deste ano. Esse processo foi repetido sequencialmente, com o modelo de 1999 sendo então ajustado para 2000, 2000 para 2001 e 2001 para 2002. Assim, cada novo modelo de cálculo dos *embeddings* acumulou seu próprio aprendizado e o aprendizado dos anos anteriores.

3.3. Extração da similaridade da palavra nova para análise temporal das variações de contexto semântico

Para o cálculo das similaridades, foram extraídas algumas palavras do vocabulário do modelo original, como símbolos especiais e *tokens* de 1 (um) caractere, referentes as letras. As palavras do vocabulário BERT foram então comparadas uma a uma com a palavra *'voip'*, gerando uma lista com todas as similaridades do vocabulário em relação à palavra nova.

Para averiguação da proximidade entre palavras foi utilizada a medida de similaridade por cossenos. Esta métrica é determinada a partir do ângulo entre dois vetores em relação à origem, de acordo com a Equação 1, que exemplifica o cálculo para dois vetores genéricos A e B, com dimensionalidade n. Dado que o valor desta equação é

normalizado, o resultado varia sempre entre -1 e 1, sendo 1 quando os vetores são idênticos, e -1 quando são opostos.

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

Esta estratégia permitiu acompanhar o processo de evolução do *embedding* da palavra de interesse, à medida que um novo ajuste era realizado e novas ocorrências da palavra eram acrescentadas.

4. DISCUSSÃO

No que tange o desempenho da obtenção de *embeddings* para palavras novas, entende-se que obter uma boa representação da palavra, implica em seu *embedding* apresentar uma alta similaridade de cossenos com *embeddings* de palavras de um mesmo contexto semântico.

Ao final de cada execução da tarefa de extração de atributos (Etapa 2) os modelos foram submetidos à análise de seus *embeddings* pelo cálculo de similaridade (Etapa 3). Essa lista com similaridades e seus respectivos *tokens* foi ordenada de forma decrescente, para destacar as palavras mais próximas da palavra nova de interesse. O resultado da primeira, terceira e última iteração podem ser observadas na Tabela 2.

Tabela 2: Resultados de similaridade obtidos após a primeira, terceira e última iteração do método

	1998		**2000**		**2002**	
	Tokens	***sim.***	***Tokens***	***sim.***	***Tokens***	***sim.***
1º	*endeavors*	.9480	*telecommunication*	.9030	*telecommunication*	.8925
2º	*gabled*	.9476	*telecommunications*	.9002	*telecommunications*	.8671
3º	*incomes*	.9468	*signalling*	.8797	*signalling*	.8621
4º	*methodist*	.9436	*ethernet*	.8755	*ieee*	.8528
5º	*stifled*	.9432	*routing*	.8667	*switching*	.8396
6º	*propped*	.9419	*integration*	.8597	*convergence*	.8393
7º	*independents*	.9409	*standardization*	.8589	*infrastructure*	.8358
8º	*impoverished*	.9403	*switching*	.8581	*upstream*	.8355
9º	*pursuits*	.9390	*infrastructure*	.8581	*standardization*	.8353

Fonte: autores.

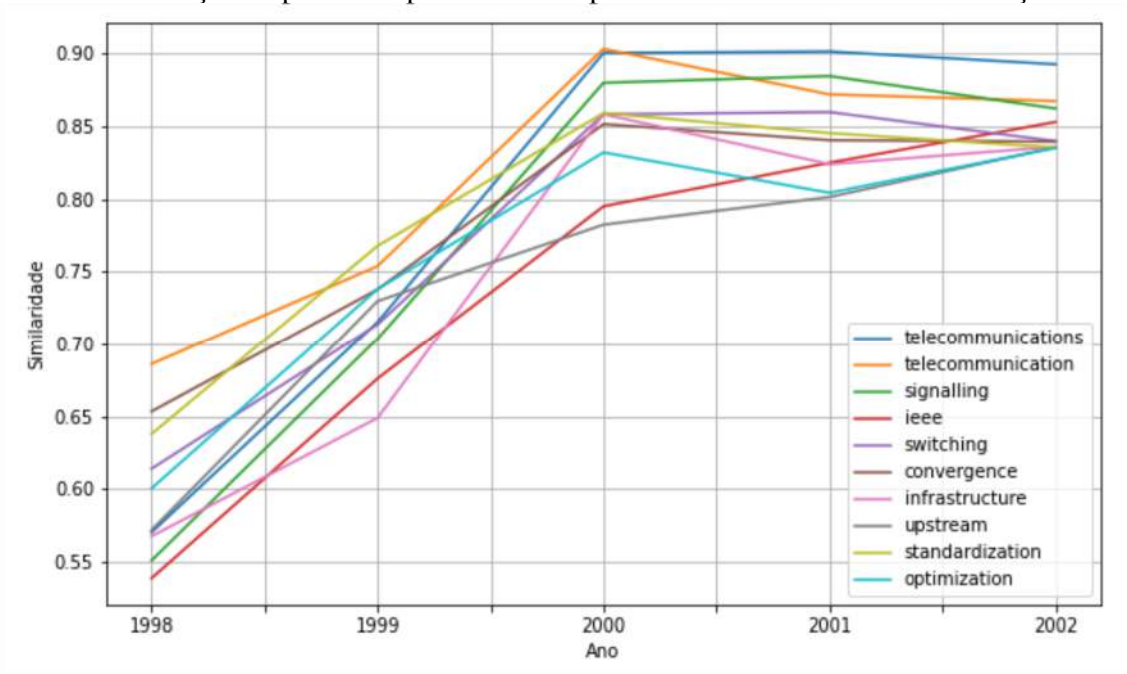
Os resultados da primeira iteração, referem-se ao *dataset* do ano de 1998; o da terceira iteração ao *dataset* do ano 2000 e a quinta, e última, iteração dizem respeito ao *dataset* de 2002. A primeira posição seria ocupada pela própria palavra do estudo, pois sendo uma palavra igual a ela mesma, a similaridade é a mais alta, igual a 1, por essa razão não foi incluída na Tabela 2, das 10 palavras mais próximas.

Para o *dataset* de 1998, embora tenham similaridades altas, todas acima de 0.93, não é possível estabelecer relações destas palavras com a palavra *'voip'*. Na ausência de relação entre as palavras consideradas próximas pela similaridade, é possível inferir que o baixo número de ocorrências da palavra '*voip'* no *dataset* do ano de 1998 não foi suficiente para produzir uma representação válida. Importante destacar que para esta primeira iteração a palavra *'voip'* foi apenas adicionada ao vocabulário BERT permitindo que o modelo reconheça a palavra como um *token* único, não sendo atribuído nenhum tipo de *embedding* significativo ou contextual antes da primeira iteração. Resultado e inferência semelhante foi obtido na iteração do método para o *dataset* para o ano de 1999.

Na terceira iteração do modelo, as similaridades mais altas começaram a apontar palavras que se encontram no mesmo campo semântico de *'voip*'. Outra inferência que pode ser obtida é a percepção de que os valores de similaridade são mais baixos do que aqueles observados anteriormente. No entanto, embora o valor das similaridades diminua, na medida em que o ajuste converge com mais iterações, a representação se aproxima de palavras inseridas no mesmo contexto.

Na quarta e quinta iteração os resultados apresentam uma tendência de estabilização. Esta tendência indica a convergência da representação da palavra nova (*embedding*), que passa a ser circundada por um grupo relativamente fixo de palavras que se encontram no mesmo campo de significância. O Gráfico 2 demonstra este processo de progressão, para as 9 palavras consideradas mais similares após a quinta iteração, ao longo dos 5 ajustes na extração de atributos do modelo. É notável a tendência de subida das similaridades ao longo dos dois primeiros ajustes, seguida de uma estabilização nos três ajustes posteriores.

Gráfico 2: Avanço temporal das palavras mais próximas exibidas na última iteração



Fonte: autores.

O modelo proposto se mostrou satisfatório, uma vez que, na medida em que os pesos do modelo são rebalanceados durante as iterações a cada *dataset*, todas as similaridades do vocabulário em relação à palavra nova são alteradas, e as palavras apontadas como mais próximas na primeira e segunda iteração, desaparecem das listas de

maiores similaridades dos anos seguintes. Ao passo que outras palavras começam a se aproximar, em uma tendência de aumento da similaridade, até se estabilizarem em um patamar fixo. Indicando uma convergência na representação da palavra nova, ou seja, seu *embedding* é capaz de representar um contexto semântico coerente.

5. CONCLUSÃO

Este trabalho contribui com um método para adaptar o modelo de linguagem BERT a um contexto específico, quando se faz necessário o *embedding* de palavras novas como *tokens* únicos, permitindo sua análise temporal. Para atingir tal objetivo realizou-se a inclusão da palavra nova de interesse no dicionário do modelo BERT pré-treinado; a extração de atributos a partir de textos em domínio específico, os quais possibilitam o ajuste dos *embeddings*, fornecendo um *embedding* coerente para a palavra alvo. Por fim, averígua-se a qualidade do *embedding* por meio de uma etapa de cálculo de similaridade entre *embeddings*. Desta forma, este trabalho contribui na resolução de tarefas de Processamento de Linguagem Natural ao indicar um caminho para a aplicação de modelos pré-treinados no processamento de palavras desconhecidas, sem a necessidade de replicação total do treinamento.

Contudo, embora os resultados atingidos no desenvolvimento do método e no estudo de caso, sejam promissores, para uma maior confiabilidade faz-se necessária a verificação do comportamento desta estratégia quando aplicada em contextos mais amplos, como na inclusão simultânea de várias palavras novas. A futura aplicação do método em outros cenários, poderá também esclarecer e determinar um número mínimo de ocorrências contextuais que uma palavra nova precisa ter, no corpus de treinamento, para se obter um *embedding* consistente. Pesquisas futuras podem explorar a aplicação do método na resolução de tarefas reais, como a identificação e classificação de neologismos. Assim como aplica-lo em análises exploratórias para investigar mudanças semânticas.

AGRADECIMENTO

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

REFERÊNCIAS

Beltagy I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing,* 3615-3620.

Bichakjian, B. H. (2017). Language evolution: How language was built and made to evolve. *Language Sciences*, 63, 119–129.

Brown, T., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.

Devlin, J., Chang, M., Lee, K, & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 1, 4171–4186.

Gontier, N. (2017). What are the levels and mechanisms/processes of language evolution? *Language Sciences*, 63, 12–43.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H.. (2022). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1), 1-23.

Hombaiah, S. A., Chen, T., Zhang, M., Bendersky, M., & Najork, M.. (2021). Dynamic language models for continuously evolving content. I*n Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2514-2524.

Hu, Z., Chen, T., Chang, K., & Sun, Y.. (2019). Few-Shot Representation Learning for Out-Of-Vocabulary Words. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4102–4112.

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E.. (2018). Diachronic word embeddings and semantic shifts: a survey. *In Proceedings of the 27th International Conference on Computational Linguistics*, 1384–1397.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J.. (2020) BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 6(4), 1234–1240.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., & Dean, J.. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111-3119.

Moon, S., & Okazaki, N.. (2020). PatchBERT: Just-in-Time, Out-of-Vocabulary Patching. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7846–7852.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S.. (2007). A Design Science Research Methodology for Information Systems Research, *Journal of Management Information Systems*, 24(3), 45-77.

Schick, T., & Schütze, H.. (2020). BERTRAM: Improved Word Embeddings Have Big Impact on Contextualized Model Performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3996–4007.

Strubell, E., Ganesh, A., & McCallum, A.. (2019). Energy and Policy Considerations for Deep Learning in NLP. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* 3645–3650.

Subakti, A., Murfi, H., & Hariadi, N.. (2022). The performance of BERT as data representation of text clustering. *Journal of big Data*, 9(1), 1-21.

Tai, W., Kung, H.T., & Dong, X.. (2020). exBERT: Extending Pre-trained Models with Domain- specific Vocabulary Under Constrained Training Resources. *Findings of the Association for Computational Linguistics*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I.. (2017). Attention is all you need. *In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, 6000-6010.

Wu, Y. et al.. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huangl, X.. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10), 1872-1897.

Zipf, G. K.. (1949). Human behavior and the principle of least effort: An introduction to human ecology, Addison-Wesley, Cambridge.