Impulsionando a descoberta de tratamentos na medicina através da representação distribuída de palavras

Matheus V. V. Berto*
Tiago A. Almeida
matheus.berto@estudante.ufscar.br
talmeida@ufscar.br
Departamento de Ciência da Computação (DComp-So), Universidade Federal de São Carlos (UFSCar)
Sorocaba, São Paulo

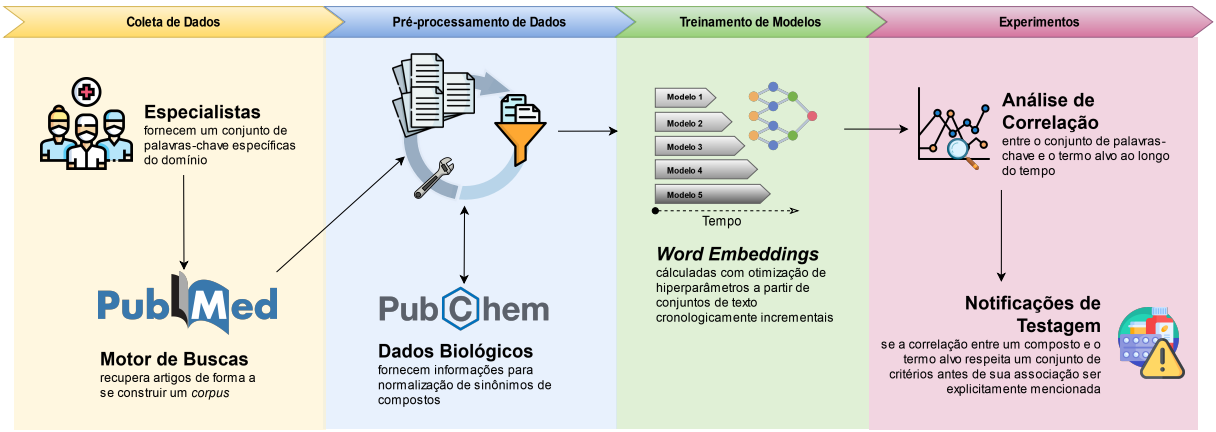

Figure 1: *Framework* proposto neste estudo. Adaptado de Berto et al. [2].

ABSTRACT

Word embeddings are mathematical and computational representations that consist of high dimensional vectors capable of encoding the meaning of terms or sentences in a text. This well-established approach enhanced many Natural Language Processing applications, since it can be easily generated from large textual datasets by a different set of algorithms. In this study, we have extended a recently discovered use of word embeddings: the ability to uncover potential implicit information in a corpus (also known as latent knowledge) that may not be achievable with human analysis alone. More specifically, our work combines word embeddings computed through diverse unsupervised methods in order to extract latent knowledge that could anticipate clinical discoveries in the field of medicine. By using a massive amount of scientific papers related to a high deadly cancer called Acute Myeloid Leukemia, our study shows that currently approved therapies could have been investigated earlier due to drug testing notifications issued by our framework. Therefore, our strategy collaborates to a faster drug analysis and biomedical discoveries. Details about our proposal and in-depth analysis of the results can be found in Berto et al. [2].

KEYWORDS

representação vetorial distribuída, embeddings de palavras, descoberta de conhecimento em bases de dados, processamento de linguagem natural, IA na medicina

1 INTRODUÇÃO

Embora dados textuais não possuam uma estrutura pré-definida, são capazes de codificar e armazenar uma imensa quantidade de informações. Devido ao crescente volume de textos acessíveis através da Internet, diversos métodos computacionais têm sido desenvolvidos com os objetivos de processar e interpretar tal tipo de conteúdo. Dentre as estratégias propostas, a representação distribuída mostrou-se altamente eficiente. Nela, cada termo do texto é transformado em um vetor de atributos, com dimensionalidade densa e previamente fixada. Tais representações vetoriais de palavras tornaram conhecidas como *word embeddings* ou *word vectors* [1]. Além disso, o uso de vetores distribuídos permite a captura de coocorrência de termos no texto, construindo um espaço vetorial que fornece aproximações coerentes de significado com relação à vizinhança (i.e., o contexto) de cada palavra [6].

Por serem vetores, as representações distribuídas geradas a partir de um *corpus* (i.e., um conjunto de documentos) podem ser manipuladas de forma a se executar operações aritméticas ou se calcular, por exemplo, distâncias vetoriais. Consequentemente, tais operações

*Matheus Berto (aluno de IC), sob a orientação do Prof. Dr. Tiago Almeida e dos colaboradores mencionados na seção de agradecimentos, realizou a pesquisa integralmente, implementou os códigos necessários, analisou os resultados e redigiu os artigos científicos publicados.

In: IV Concurso de Trabalhos de Iniciação Científica (CTIC 2024). Anais Estendidos do XXX Simpósio Brasileiro de Sistemas Multimídia e Web (CTIC'2024). Juiz de Fora/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2024.

ISSN 2596-1683

possibilitam a descoberta de relações – explícitas ou implícitas – entre os textos [5]. Assim, termos próximos no espaço vetorial tendem a compartilhar características sintáticas ou semânticas.

1.1 Motivação

O presente estudo foi motivado por recentes demonstrações de uso de *word vectors* para extração de conhecimento latente codificado em documentos da literatura científica. Mais precisamente, um dos trabalhos mais notórios sobre o tema, conduzido por Tshitoyan et al. [8], provou ser possível empregar *word embeddings* de forma a identificar relações complexas entre materiais termoelétricos e antecipar sua descoberta através de análises temporais de modelos não supervisionados. A avaliação de tal estratégia sobre outras áreas de conhecimento também foi sugerida pelos autores.

Seguindo a recomendação de Tshitoyan et al. [8], Shetty and Ramprasad [7] e Yang [9] também propuseram protocolos para avaliação de inferência de informações sobre outros temas. Entretanto, nenhuma dessas pesquisas abordou textos relacionados à saúde humana. O aprimoramento da busca por descobertas científicas nessa área é de considerável importância e relevância.

1.2 Objetivos e contribuições

Com base nos estudos de Tshitoyan et al. [8], Shetty and Ramprasad [7] e Yang [9], este trabalho tem por objetivo avaliar a possibilidade de decodificação de conhecimento latente na literatura biomédica. Mais precisamente, a pesquisa busca analisar se o emprego de *word embeddings* é capaz de acelerar descobertas acerca de tratamentos atualmente aprovados para a Leucemia Mieloide Aguda (LMA ou, em inglês, AML) – uma forma rara e altamente letal de câncer [4].

O restante deste artigo apresenta uma breve revisão da literatura sobre o tema e indica a metodologia empregada no estudo, além dos resultados obtidos. Entretanto, este texto representa uma versão sumarizada do texto completo do estudo, redigido exclusivamente para participação no CTIC/WebMedia 2024. O artigo que descreve em detalhes toda a pesquisa desenvolvida foi publicado pelos autores e está disponível em Berto et al. [2].

2 TRABALHOS RELACIONADOS

Dentre os primeiros métodos para representação computacional de textos, a codificação *one-hot* constitui-se em transformar termos de um documento em vetores esparsos. Estes vetores possuem dimensionalidade igual ao tamanho total de palavras presentes no vocabulário do *corpus* (V). Todos os elementos dos vetores são preenchidos com zeros, exceto aquele o qual corresponde ao índice do termo em questão no vocabulário. Entretanto, essa estratégia apresenta alto custo computacional, visto que quanto maior a extensão do vocabulário, mais esparsos os vetores se tornam, dificultando operações aritméticas e exigindo mais uso de memória.

Por outro lado, *word embeddings* são capazes de representar coerentemente características semânticas e sintáticas das palavras, além de não apresentarem dimensionalidade obrigatoriamente equivalente à quantidade de termos distintos presentes nos documentos. Um dos trabalhos pioneiros responsáveis pela popularização da representação distribuída foi proposto por Mikolov et al. [5]. Nele, um novo algoritmo denominado *Word2Vec* foi implementado através de duas arquiteturas distintas: *Continuous Bag-of-Words* (CBOW) e *Skip-Gram*. Ambos os métodos empregam redes neurais rasas com o objetivo de predizer termos próximos vinculados ao contexto de uma palavra-alvo ou vice-versa.

A Figura 2 ilustra o funcionamento da arquitetura *Skip-Gram*. A partir da codificação *one-hot* inicial de um termo na camada de entrada, cálculos são executados pelos neurônios da camada oculta. Finalmente, na camada de saída, uma função exponencial normalizada (ou *softmax*) gera um vetor, no qual os elementos indicam a probabilidade de cada termo de V pertencer à vizinhança da palavra de entrada.

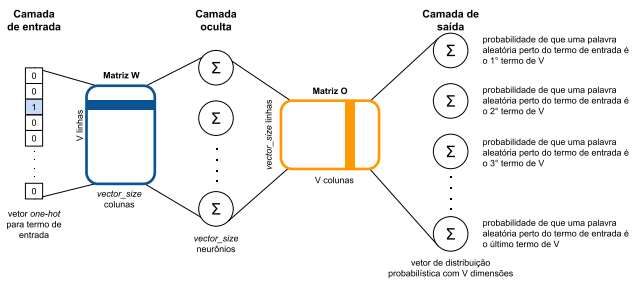

Figure 2: Arquitetura *Skip-Gram*. Adaptado de Berto et al. [2].

O vetor normalizado de probabilidades gerado pela camada de saída da arquitetura não é o artefato mais útil da arquitetura. Na verdade, são duas matrizes internas – W e O – populadas durante o treinamento dos modelos que armazenam as representações distribuídas. As linhas da matriz W são as *word embeddings*, enquanto as colunas da matriz O são chamadas de *output embeddings*. As matrizes possuem uma quantidade V de linhas e colunas, respectivamente. O produto escalar entre dois vetores, sendo cada um deles pertencente a uma dessas matrizes, expressa a probabilidade de ambos os termos coocorrerem em um mesmo texto (Figura 3). Dessa forma, o valor do produto escalar é diretamente proporcional ao nível de correlação entre as palavras analisadas.

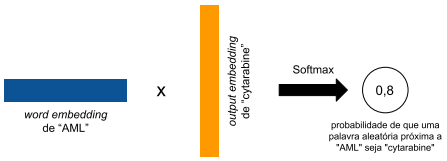

Figure 3: Cálculo de probabilidade de contexto. Adaptado de Berto et al. [2].

Outro método popular para se computar *word vectors* é o *FastText* [3]. Através dele, além de vetores para cada palavra de um vocabulário V, também são gerados vetores de sub-palavras, representadas por n-gramas (i.e., subconjuntos de n caracteres contíguos em um texto). Tal método é capaz de melhorar a interpretação subjetiva codificada a partir dos vetores e representar termos pouco frequentes ou com estrutura complexa.

Como uma aplicação refinada dos algoritmos e arquiteturas de representação distribuída acima mencionados, Tshitoyan et al. [8] empregaram modelos baseados em *Word2Vec* e *Skip-Gram* gerados de forma cronológica e incremental capazes de predizer elementos da ciência dos materiais atualmente conhecidos. Semelhantemente, Yang [9] também demonstraram a possibilidade de codificar conhecimento em textos relacionados a polímeros. A partir do processamento de 500.000 documentos, o estudo foi capaz de identificar os principais materiais citados pela literatura ao longo dos anos.

Yang [9] também demonstrou a alta capacidade de usar *word embeddings* para realizar predições significativas relacionadas a elementos sobre os estudos de células eletroquímicas de combustão. Dessa forma, com base nos resultados recentes publicados na literatura, este trabalho estende as estratégias empregadas de forma a expandir a análise de conhecimento latente no contexto da LMA.

3 METODOLOGIA

Esta seção contempla as principais etapas executadas na realização do presente estudo. Elas estão sumarizadas na Figura 1.

3.1 Coleta e pré-processamento de dados

A partir de um conjunto pré-definido de palavras-chave escolhidas por um especialista em LMA, somado a nomes e sinônimos de compostos atualmente aprovados para o tratamento da doença[1], cerca de 272.000 artigos científicos foram coletados. Todos os artigos recuperados estão na língua inglesa e tiveram sua coleta ou filtragem através de *scripts* escritos na linguagem Python com acesso ao motor de buscas PubMed[2].

O conjunto de textos coletados passou então por uma etapa de pré-processamento, na qual sinônimos de drogas e outras substâncias relevantes ao projeto foram normalizadas para um único termo comum. A substituição automática foi realizada a partir da consulta a arquivos contendo milhões de registros biológicos disponibilizados pela plataforma PubChem [3] e teve por objetivo mitigar a sensibilidade dos algoritmos à coocorrência de termos no texto. Sinônimos da LMA também foram todos transformados para a abreviação universal "AML". Por fim, outras técnicas, como o tratamento de caracteres especiais, palavras irrelevantes (*stopwords*) e lematização foram aplicadas utilizando-se o conjunto de bibliotecas computacionais NLTK [4].

3.2 Modelos de representação distribuída

Os algoritmos *Word2Vec* e *FastText* foram executados sobre conjuntos de documentos cronologicamente ordenados. O *corpus* abrangeu artigos escritos entre os anos 1963 e 2022, totalizando 60 modelos de representação distribuída que agregam textos publicados antes ou durante seu respectivo ano. As implementações dos algoritmos mencionados foram realizados com o uso da biblioteca *Gensim* [5].

3.3 Análise de conhecimento latente

A partir dos artigos coletados da literatura, foi construída uma linha do tempo responsável por registrar a primeira ocorrência de cada uma das drogas-alvo deste estudo em publicações diretamente ligadas ao contexto da LMA.

A evolução histórica da correlação entre cada uma das drogas-alvo e o termo "AML", calculada a partir do produto escalar (Figura 3), foi analisada anualmente. Períodos de tamanho fixo w que apresentam um crescimento significativo na correlação entre os termos são de especial interesse. Tal aumento expressivo indica que o composto em análise deve ser recomendado para investigação por especialistas em LMA.

Os compostos avaliados em cada intervalo precisam atender critérios rígidos para serem considerados significativos. Caso um composto respeite algum dos critérios estabelecidos, o *framework* proposto emite uma notificação de testagem no ano seguinte à tal identificação.

(1) a derivada da curva de produto escalar é positiva e a taxa de crescimento é maior ou igual a valor mínimo x; ou
(2) o aumento final do produto escalar durante o período, calculado pela subtração entre o último e o primeiro valor registrado, deve ser igual ou superior a um segundo parâmetro t, sendo $t >> x$.

4 RESULTADOS

Diferentes valores para os parâmetros w, x e t envolvidos nos critérios de análise foram testados empiricamente com ajuda de especialistas, sendo {3; 2; 4} a combinação final utilizada. A partir da aplicação de tais parâmetros ao protocolo de validação descrito na seção anterior sobre os históricos de produto escalar de cada um dos compostos-alvo, o *framework* construído emitiu notificações indicadas na Figura 5.

O sistema foi capaz de emitir 11 notificações para 6 drogas distintas, sendo *Azacitidine* e *Prednisone* as mais frequentes (4 e 3 vezes, respectivamente). Considerando apenas a primeira notificação para cada droga, o *framework* proposto sugeriu a testagem com cerca de 5 anos antes da real data de associação entre o composto e a doença. Notavelmente, o composto *Arsenic Trioxide* apresentou maior antecedência em seu valor de correlação, sendo sugerido para testagem 11 anos antes da sua primeira associação com LMA.

Também foi analisado o real impacto do uso do *framework* proposto, i.e., foi quantificado o quanto as notificações de testagem emitidas seriam capazes de acelerar as descobertas. Iniciando em 1963, a estratégia de ranqueamento por produto escalar foi empregada. Para cada ano, foram selecionados apenas os três primeiros compostos que apresentaram maior correlação com LMA. Assim, apenas as melhores predições dos modelos foram consideradas. Conforme indicado na Figura 4, o uso dessa estratégia (curva de cor laranja) foi capaz de acelerar a porcentagem de descoberta dos 21 compostos-alvo deste estudo em até 2.3× após 1968, quando comparado ao cenário sem o emprego do *framework* proposto.

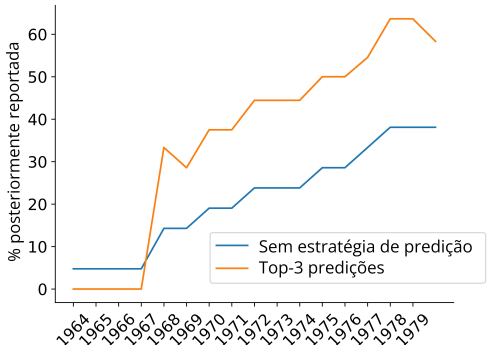

Figure 4: Porcentagem de compostos preditos pelos modelos e que foram posteriormente associados à LMA. A linha laranja compara o uso de apenas as três primeiras predições anuais a uma escolha aleatória de drogas. Adaptado de Berto et al. [2].

[1] https://www.cancer.gov/about-cancer/treatment/drugs/leukemia#3
[2] https://pubmed.ncbi.nlm.nih.gov/
[3] https://pubchem.ncbi.nlm.nih.gov/
[4] https://www.nltk.org/
[5] https://radimrehurek.com/gensim_3.8.3/

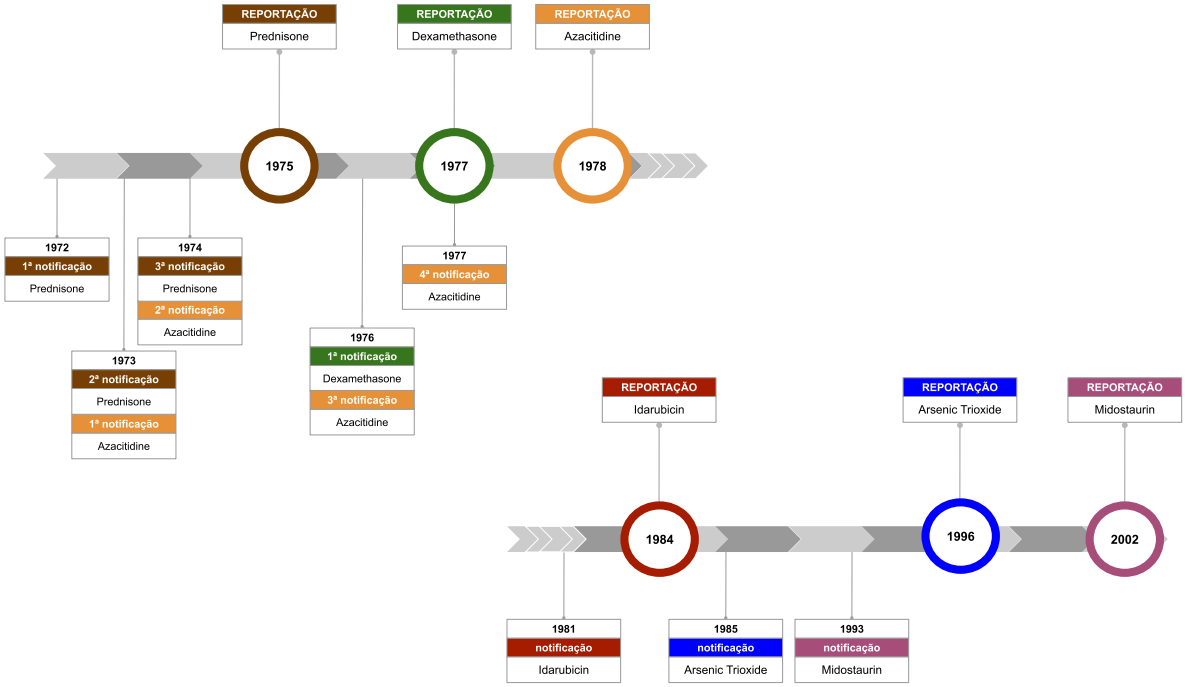

Figure 5: Notificações de testagem de compostos emitidas pelo sistema proposto. Adaptado de Berto et al. [2].

5 CONCLUSÕES E TRABALHOS FUTUROS

O presente artigo apresentou um resumo da proposta detalhada em Berto et al. [2]: um sistema baseado em modelos de representação distribuída capazes de codificar coerentemente informações relacionadas à saúde humana, especificamente sobre a Leucemia Mieloide Aguda (LMA). O estudo foi baseado em diversas pesquisas da literatura científica que demonstraram êxito de suas aplicações em áreas de conhecimento distintas. Portanto, a extensão de tais pesquisas sobre o contexto da LMA apresenta originalidade e embasamento técnico.

Os resultados obtidos comprovaram que o *framework* proposto foi capaz de orientar investigações de forma a potencialmente impulsionar o descobrimento de tratamentos para a referida doença. A estratégia desenvolvida apresenta vantagens como o emprego de algoritmos consolidados e computacionalmente baratos, exigindo pouco recurso computacional para sua execução [2].

Apesar dos resultados encontrados serem bastante promissores, o estudo descrito possui certas limitações. As representações distribuídas geradas pelos algoritmos *Word2Vec* e *FastText* são estáticas, i.e., uma determinada palavra do *corpus* será sempre vinculada ao mesmo vetor, independentemente do contexto da sentença no qual ela esteja presente. Portanto, a adaptação do sistema para o emprego de representações contextuais consideradas estado-da-arte em diversas aplicações pode potencializar ainda mais os resultados.

AGRADECIMENTOS

Este trabalho foi financiado pelas agências de fomento FAPESP (2021/13054-8 e 2022/07236-9), Capes e CNPq. Os autores agradecem a estreita e inestimável colaboração de Breno L. Freitas (Shopify Inc.), Profa. Dra. Carolina Scarton (The University of Sheffield) e Prof. Dr. João A. Machado-Neto (ICB/USP) nesta pesquisa. Os autores também são gratos à Priscila Portela Costa pela sua contribuição durante o planejamento inicial do projeto.

REFERENCES

[1] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A Neural Probabilistic Language Model. *JMLR* 3 (2003), 1137–1155. https://doi.org/10.1162/153244303322533223

[2] Matheus V. V. Berto, Breno L. Freitas, Carolina Scarton, João A. Machado-Neto, and Tiago A. Almeida. 2024. Accelerating discoveries in medicine using distributed vector representations of words. *Expert Systems with Applications* 250 (2024), 123566. https://doi.org/10.1016/j.eswa.2024.123566

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *Trans. of the ACL* 5 (07 2016), 12. https://doi.org/10.1162/tacl_a_00051

[4] Bob Löwenberg, Gert J. Ossenkoppele, Wim van Putten, Harry C. Schouten, Carlos Graux, Augustin Ferrant, Pieter Sonneveld, Johan Maertens, Mojca Jongen-Lavrencic, Marie von Lilienfeld-Toal, Bart J. Biemond, Edo Vellenga, Marinus van Marwijk Kooy, Leo F. Verdonck, Joachim Beck, Hartmut Döhner, Alois Gratwohl, Thomas Pabst, and Gregor Verhoef. 2009. High-Dose Daunorubicin in Older Patients with Acute Myeloid Leukemia. *New England J. of Medicine* 361, 13 (Sept. 2009), 1235–1248. https://doi.org/10.1056/nejmoa0901409

[5] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781

[6] Magnus Sahlgren. 2008. The distributional hypothesis. *Italian J. of Linguistics* 20 (01 2008), 33–54.

[7] Pranav Shetty and Rampi Ramprasad. 2021. Automated knowledge extraction from polymer literature using natural language processing. *iScience* 24, 1 (Jan. 2021), 101922. https://doi.org/10.1016/j.isci.2020.101922

[8] Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander R Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* 571 (July 2019), 95–98. https://doi.org/10.1038/s41586-019-1335-8

[9] Feifan Yang. 2022. Natural Language Processing Applied on Large Scale Data Extraction from Scientific Papers in Fuel Cells. In *Proc. of the 5th NLPIR* (Sanya, China) *(NLPIR 2021)*. ACM, New York, NY, USA, 168–175. https://doi.org/10.1145/3508230.3508256