Um Estudo Aprofundado sobre Grupos Semânticos de Palavras - CluWords - em tarefas de PLN

Felipe Viegas
frviegas@dcc.ufmg.br
UFMG - Minas Gerais - Brasil

Leonardo Rocha
lcrocha@ufsj.edu.br
UFSJ - Minas Gerais - Brasil

Marcos André Gonçalves
mgoncalv@dcc.ufmg.br
UFMG - Minas Gerais - Brasil

ABSTRACT

This Ph.D. dissertation focused on proposing, designing and evaluating a novel textual document representation that exploits the "best of two worlds": efficient and effective frequentist information (TFIDF representations) with semantic information derived from word embedding representations. In more details, our proposal – called **CluWords** – groups syntactically and semantically related words into clusters and applies domain-specific and application-oriented filtering and weighting schemes over them to build powerful document representations especially tuned for the task in hand. We apply our novel Cluword concept to four Natural Language Processing (NLP) applications, related to topics from WebMedia interest: topic modeling, hierarchical topic modeling, sentiment lexicon building, and sentiment analysis. Some of the novel contributions of this dissertation include: (i) the introduction of a new data representation; (ii) the design of CluWords' components capable of improving the effectiveness of Topic Modeling, Hierarchical Topic Modeling and Sentiment Analysis applications; (iii) the proposal of two new topic quality metrics to assess the topical quality of the hierarchical structures. Our extensive experimentation demonstrates that CluWords produce the current state-of-the-art topic modeling and hierarchical topic modeling. For sentiment analysis, our experiments show that CluWords filtering and weighting can mitigate semantic noise, surpassing powerful Transformer architectures in the task. Our results were published in some of the most important conferences in journals of the field, as detailed in this document. Our work was supported by two *Google Research Awards*.

KEYWORDS

Representação de dados, modelagem de tópicos, análise de sentimento, processamento de linguagem natural.

1 INTRODUÇÃO

1.1 Contexto e problema

Enquanto os mundos acadêmico e industrial correm a passos largos atrás de (Grandes) Modelos de Linguagem (aka Large Language Models [4]) cada vez mais complexos, caros de serem treinados, e difíceis de interpretar, essa tese apresenta soluções simples, mas elegantes, baseadas em engenharia de dados para modelagem de texto que tratam problemas relacionados a ruído e escassez de informações em documentos. As soluções propostas são competitivas (ou até mesmo superiores) com o estado-da-arte em termos de efetividade, em aplicações tais como Modelagem de Tópicos e Análise de Sentimentos, sendo ao mesmo tempo bastante eficientes e tendo uma alta capacidade de interpretação (explicabilidade).

A contribuição central da presente tese é um conceito inovador na área de PLN denominado Cluwords – uma nova representação textual que aproveita a eficiência e a interpretabilidade das representações (matriciais) tradicionais baseadas em frequências de palavras (e.g., TFIDF), ao mesmo tempo que explora as capacidades semânticas de modelos modernos baseados em embeddings[1] de palavra.

1.2 Aderência ao WebMedia

Do ponto de vista técnico, a aderência da presente tese ao WebMedia começa pelo próprio título no qual, explicitamente, mencionamos um de seus tópicos de interesse: Processamento de Linguagem Natural. Nossa proposta de representação semântica de texto - CLuwords - foi instanciada em três conjuntos de aplicações: Modelagem de Tópicos, Modelagem de Tópicos Hierárquica e Análise de Sentimento. Todas essas aplicações estão diretamente relacionadas a dois tópicos de interesse do Webmedia: 1) Engenharia de Documentos, Modelos e Linguagens e 2) IA, Aprendizado de Máquina e Aprendizado Profundo. Por fim, ainda do ponto de vista técnico, conforme detalhado nos subprodutos, a presente tese resultou na publicação de cinco artigos no WebMedia (três em 2023, 1 em 2022 e 1 em 2021), correspondendo a um fato concreto que a mesma trata de um problema de claro interesse para as áreas de Sistemas Multimídia e Web. Do ponto de vista qualitativo, uma representação de texto semântica de baixo custo computacional, porém efetiva e competitiva com representações caras e complexas, é de interesse para uma vasta gama de aplicações relacionadas a Web, tais como máquinas de busca, recuperação de informação, análise de conteúdo, dentre outras. Atualmente, essa representação vem sendo utilizada em tarefas de expansão de consultas em uma colaboração internacional com a Universit'a di Padova.

2 METODOLOGIA

2.1 Solução

A estrutura CluWords compreende três etapas fundamentais — agrupamento, filtragem e ponderação — destinadas a construir uma representação mais informativa para coleções de dados textuais adaptadas a cada cenário de aplicação. CluWords envolve grupos (clusters) de embeddings de palavras semanticamente relacionados, formados por meio da aplicação de funções de distância e mecanismos de filtragem personalizáveis. As CluWords procuram explorar as similaridades sintáticas e semânticas de embeddings de palavras,

[1] Representações vetoriais de palavras de alta dimensionalidade, cuja posição espacial é correlacionada com sua semântica (em relação a outras palavras que ocorrem em contextos textuais similares).

In: VI Concurso de Teses e Dissertações (CTD 2024). Anais Estendidos do XXX Simpósio Brasileiro de Sistemas Multimídia e Web (CTD'2024). Juiz de Fora/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2024.

ISSN 2596-1683

acoplando aos clusters filtros para o tratamento de ruído e para ponderação dos termos de maneira adequada, de forma a construir representações de palavras enriquecidas e adaptáveis à tarefa alvo.

2.2 Objetivo

O principal objetivo da tese é fornecer evidências para "comprovar" a hipótese de que a Cluwords é uma alternativa melhor (mais eficaz, eficiente e interpretável) para representar textos, especialmente em conjuntos de dados pequenos, mais "ruidosos" e que sofrem com a escassez de informações, pois capturam relações semânticas junto com informações frequentistas, cruciais para tarefas de PLN.

As principais questões de pesquisa, derivadas da hipótese foram: (i) As CluWords podem ser efetivamente exploradas para avançar o estado-da-arte em tarefas de PLN e recuperação de informações? (ii) Mecanismos de filtragem e ponderação específicos para certas tarefas seriam capazes de efetivamente adaptar as CluWords a diferentes cenários de PNL?. Questões específicas de pesquisa, considerando três cenários de aplicação, incluem:

- Modelagem de Tópicos (MT): (i) Podemos explorar as CluWords para melhorar a representação de documentos para modelagem de tópicos? (ii) As CluWords podem adicionar mais informações aos modelos hierárquicos de modelagem de tópicos em níveis mais profundos da hierarquia?
- Análise de Sentimentos (AS): (i) As CluWords podem ser usadas para superar problemas de falta de informação em tarefas de análise de sentimento? (ii) A polaridade/intensidade e a classe gramatical (PoS) podem ser usadas para filtrar palavras das CluWords para análise de sentimento?

2.3 Avaliação

As análises experimentais forneceram evidências para responder positivamente à primeira questão de pesquisa no contexto da MT. Por meio de experimentos com 12 conjuntos de dados e oito linhas de base, confirmou-se que as CluWords constroem tópicos melhores e enriquecem significativamente as representações de documentos.

Em relação à MT Hierárquica (MTH), a tese apresenta um novo método não-probabilístico denominado CluHTM, que explora a informação semântica global fornecida pela representação CluWords e uma aplicação original de uma medida de estabilidade para definir a "forma" da hierarquia. São apresentadas duas variantes do método CluHTM, uma que explora embeddings estáticos (f-CluHTM) e outra que usa dinâmicos (c-CluHTM). Ambas as variantes CluHTM se destacaram, sendo cerca de duas vezes mais eficazes que as linhas de base do estado-da-arte. Até o presente momento (2024) nenhuma abordagem conhecida superou nossos resultados.

A tese ainda propôs, como contribuição adicional, novas métricas para avaliar métodos MTH. As métricas de qualidade propostas avaliam aspectos relacionados à consistência topológica e à estrutura semântica hierárquica que são importantes para métodos hierárquicos. Esses são aspectos diferentes e complementares daqueles capturados por métricas tradicionais de MT, como NPMI e Coerência. Em outras palavras, as novas métricas de qualidade de tópicos capturam comportamentos distintos dos tópicos construídos, incluindo duplicidade e consistência semântica. Os resultados experimentais mostram que novos métodos c-CluHTM e f-CluHTM apresentam os melhores resultados na construção de uma estrutura hierárquica quando comparados com o estado-da-arte.

Fazendo a transição para o domínio da Análise de Sentimentos (AS), em relação às questões de pesquisa RQ2.i e 2.ii, a tese primeiramente fornece hipóteses formais apoiadas por fortes evidências empíricas e experimentais que demonstram o potencial de exploração de CluWords em AS. Além disso, é proposta uma técnica nova, simples, mas muito eficaz, para expandir léxicos humanamente construídos. O método proposto pode usar a representação geral fornecida por embeddings de palavras e seus relacionamentos (capturados por simples cálculos de distância) para produzir léxicos de alta cobertura que melhoram significativamente a precisão dos métodos de AS. Complementarmente apresentamos uma nova instanciação da CluWords para SA – CluSent – que explora a expansão semântica e aborda problemas de escassez de informação e ruído. A representação CluSent é construída por um pipeline dinâmico de instanciações para construir representações de documentos adaptadas às características dos conjuntos de dados. A avaliação experimental revela que o CluSent, por meio de filtragem baseada em Part-of-Speech e ponderação de sentimento (i.e., polaridade), é tão eficaz quanto os melhores métodos Transformers de última geração para a tarefa de AS.

3 AVANÇO NO ESTADO-DA-ARTE

A solução proposta para Modelagem de Tópicos (MT) e MT Hierárquica (MTH) são o estado-da-arte, superando estratégias eficazes tais como, BERTopic [2]. Até o presente momento os resultados reportados na tese não foram superados por nenhuma outra estratégia da literatura. A solução de expansão de léxicos também superou estratégias não supervisionadas, tais como VADER [3]. Em AS o CluSent se equiparou com estratégias complexas e caras, tais como BERT [1], sendo amplamente mais explicável e eficiente.

4 CONTRIBUIÇÕES

O trabalho resultou na publicação de **oito** produções diretamente da tese, sendo 4 em conferências (CIKM - A1; WSDM - A1, ACL - A1; e WebMedia - A4) e quatro em periódicos (Information Systems -A2, Scientometrics -A1, Computational Linguistic - A1; e Journal on Interactive Systems - B1). Outros **16 artigos** foram publicados tendo o doutorando como coautor, em temas relacionados à sua tese (nove A1, um A2, um A3 e cinco A4). A tese de doutorado recebeu dois *Google Latin America Research Awards (LARA)* – oitava e nona edições.

AGRADECIMENTOS

Este trabalho foi financiado por CNPq, CAPES, Google, Fapemig e AWS

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018). https://arxiv.org/abs/1810.04805

[2] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[3] Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media.*

[4] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. 18, 6 (2024). https://doi.org/10.1145/3649506