On the Role of Semantic Word Clusters — CluWords — in Natural Language Processing (NLP) Tasks

Felipe Viegas[1], **Leonardo Rocha**[2], **Marcos André Gonçalves**[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)

[2]Departamento de Ciência da Computação – Universidade Federal de São João Del Rei (UFSJ)

{frviegas,mgoncalv}@dcc.ufmg.br, lcrocha@ufsj.edu.br

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 detail, 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 at hand. We apply our novel Cluwords concept to four Natural Language Processing (NLP) applications: 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 produces 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.*

Resumo. *Esta tese de doutorado tem como foco a proposta, concepção e avaliação de uma nova representação textual de documentos que combina o "melhor de dois mundos": a informação frequentista, eficiente e eficaz (representações TFIDF), com informações semânticas derivadas de representações de word embeddings. Especificamente, nossa proposta — denominada* ***CluWords*** *— agrupa palavras relacionadas sintática e semanticamente em clusters e aplica esquemas de filtragem e ponderação específicos para o domínio e orientados à aplicação, visando construir representações documentais poderosas e ajustadas às necessidades específicas de cada tarefa. O conceito inovador de CluWords foi aplicado em quatro aplicações de Processamento de Linguagem Natural (PLN): modelagem de tópicos, modelagem hierárquica de tópicos, construção de léxicos de sentimentos e análise de sentimentos. As contribuições principais desta dissertação incluem: (i) a introdução*

de uma nova representação de dados; (ii) o desenvolvimento de componentes do CluWords capazes de aprimorar a eficácia em aplicações de Modelagem de Tópicos, Modelagem Hierárquica de Tópicos e Análise de Sentimentos; (iii) a proposta de duas novas métricas para avaliar a qualidade tópica das estruturas hierárquicas. Nossos extensos experimentos demonstram que CluWords alcança o estado da arte atual em modelagem de tópicos e modelagem hierárquica de tópicos. No contexto da análise de sentimentos, os resultados mostram que a filtragem e ponderação proporcionadas pelo CluWords podem mitigar o ruído semântico, superando até mesmo arquiteturas poderosas baseadas em Transformadores. Os resultados foram publicados em algumas das principais conferências e revistas científicas da área, conforme detalhado neste documento. Este trabalho foi apoiado por dois Google Research Awards.

1. Dados da Tese Defendida

- Data da Defesa: 10/07/2023
- Programa de Pos-Graduação: PPGCC - Universidade Federal de Minas Gerais
- Categoria: doutorado
- Membros da Banca: Prof. Pedro Olmo Stancioli Vaz de Melo – UFMG, Prof. Rodrygo Luis Teodoro Santos – UFMG, Profa. Viviane Pereira Moreira – UFRGS, Profa. Renata Vieira – Universidade de Évora

2. Contexto e problema

Enquanto os mundos acadêmico e industrial correm a passos largos atrás de (Grandes) Modelos de Linguagem (aka Large Language Models [Yang et al. 2024]) cada vez mais complexos, caros[1] de serem treinados, e difíceis de entender e interpretar[2], essa tese apresenta soluções simples, mas elegantes, baseadas em engenharia de dados para modelagem de documentos textuais que possam tratar 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).

As relevância e importância teórica e prática da tese são numerosas e demonstram que persistência e criatividade permitem "remar contra o corrente" e ainda assim obter resultados científicos e práticos de alto impacto e relevância. De fato, as contribuições da tese vão além de questões técnicas e científicas, sendo uma demonstração prática e efetiva sobre como pequenos e médios laboratórios de pesquisa em universidades, e pequenas e médias empresas do ramo, com criatividade e persistência, podem ainda ter impacto em áreas de pesquisa como Processamento de Linguagem Natural (PLN) e Inteligência Artificial (IA), hoje dominadas pelos big players.

[1]Tanto do ponto de vista econômico quanto do impacto ambiental em termos de eletricidade e emissão de carbono.

[2]Vide o AI Index Report do Stanford, recenmente publicado: `https://aiindex.stanford.edu/wp-content/uploads/2024/04/HAI_AI-Index-Report-2024.pdf`

3. Objetivo

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

As principais questões (gerais) de pesquisa, derivadas da hipótese principal, 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, derivadas das questões gerais, 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?

4. Contribuição

A contribuição central da tese de Felipe Viegas 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[3] de palavra.

5. Resumo da 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 cenários de aplicação específicos. CluWords envolvem 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, acoplando aos clusters filtros para o tratamento de ruído[4] e para ponderação dos termos de maneira adequada, de forma a construir representações de palavras enriquecidas e adaptáveis à tarefa alvo.

[3]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).

[4]Definido no contexto como sinais textuais (e.g., palavras) e linguísticos (e.g., polaridade, classe gramatical) que têm a capacidade de reduzir a eficácia da representação para uma tarefa específica.

6. Avanço no estado-da-arte

A solução proposta para Modelagem de Tópicos (MT) e MT Hierárquica (MTH) são o estado-da-arte, superando estratégias populares e eficazes tais como, SeaNMF [Shi et al. 2018], GPU-DMM [Li et al. 2017], BERTopic [Grootendorst 2022] e HLDA [Griffiths et al. 2004]. 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 [Hutto and Gilbert 2014] e SENTPRO [Hamilton et al. 2016]. Em Análise de Sentimentos o CluSent se equiparou com estratégias complexas e caras, tais como BERT [Devlin et al. 2018] e L_MIXED [Sachan et al. 2019], sendo amplamente mais explicável e eficiente.

7. Avaliação

As análises experimentais forneceram evidências convincentes 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 podem construir tópicos melhores e enriquecer significativamente as representações de documentos.

Aprofundando-se na segunda questão de pesquisa, que trata de MT Hierárquica (MTH), a tese apresenta um novo método não-probabilístico não supervisionado denominado CluHTM. Este método 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 de tópico propostas avaliam aspectos relacionados à consistência topológica (ou redundância) 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, que não consideram relações topológicas entre tópicos. Em outras palavras, as novas métricas de qualidade de tópicos capturam comportamentos distintos dos tópicos construídos, incluindo duplicidade e consistência semântica. Os resultados experimentais também mostram que novos métodos c-CluHTM e f-CluHTM apresentam os melhores resultados na construção de uma estrutura hierárquica evitando redundância 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. Complemen-

tarmente 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.

8. Produção Científica, Técnica e Premiações

Publicações diretas da Tese:

1. **Felipe Viegas**, S. D. Canuto, W. Cunha, C. França, C. Valiense, L. Rocha, M. A. Gonçalves: CluSent - Combining Semantic Expansion and De-Noising for Dataset-Oriented Sentiment Analysis of Short Texts. WebMedia 2023: 110-118 (Qualis A4; h5-index: 7)
2. **Felipe Viegas**, A. Júnior, P. Cecilio, E. Tuler, W. Meira Jr., M. A. Gonçalves, L. Rocha: Semantic Academic Profiler (SAP): a framework for researcher assessment based on semantic topic modeling. Scientometrics 127(8): 5005-5026 (2022) (Qualis A1; h5-index: 69, Impact Factor: 3.9)
3. **Felipe Viegas**, W. Cunha, C. Gomes, A. Souza Jr, L. Rocha, M. A. Gonçalves: CluHTM - Semantic Hierarchical Topic Modeling based on CluWords. AAAI Annual Meeting of the Association for Computational Linguistics (ACL) 2020 : 8138-8150. (Qualis A1, h5-index: 192)
4. **Felipe Viegas**, M. S. Alvim, S. D. Canuto, T. Rosa, M. A. Gonçalves, L. Rocha: Exploiting semantic relationships for unsupervised expansion of sentiment lexicons. Information Systems. 94: 101606 (2020) (Qualis A2; h5-index: 42, Impact Factor: 3.7)
5. **Felipe Viegas**, S. D. Canuto, C. Gomes, W. Luiz, T. Rosa, S. Ribas, L. Rocha, M. A: Gonçalves: CluWords: Exploiting Semantic Word Clustering Representation for Enhanced Topic Modeling. ACM International Conference on Web Search and Data Mining (WSDM) 2019: 753-761(Qualis A1; h5-index: 75)
6. **Felipe Viegas**, W. Luiz, C. Gomes, A. Khatibi, S. D. Canuto, F Mourão, T. Salles, L. Rocha, M. A. Gonçalves: Semantically-Enhanced Topic Modeling. ACM Conference on Information and Knowledge Management (CIKM) 2018 893-902. (Qualis A1; h5-index: 79)

Artigos Submetidos:

1. **Felipe Viegas**, W. Cunha, A. Pereira, C. França, C. Andrade, E. Tuler, L. Rocha, M. A. Gonçalves. Exploiting Contextual Embeddings in Hierarchical Topic Modeling and Investigating the Limits of the Current Evaluation Metrics. Computational Linguistics (2024)(2o. Round of review) (h5-index: 34, Impact Factor: 9.3)

Publicações inspiradas ou influenciadas pela tese:

1. W. Cunha, **Felipe Viegas**, C. França, T. Rosa, L. Rocha, M. A. Gonçalves: A Comparative Survey of Instance Selection Methods applied to Non-Neural and Transformer-Based Text Classification. ACM Computing Surveys. 55(13s): 265:1-265:52 (2023) (7) (Qualis A1; h5-index: 190; Impact Factor: 16.2).

2. C. Valiense, F. Belém, W. Cunha, C. França, **Felipe Viegas**, L. Rocha, M. A. Gonçalves: On the class separability of contextual embeddings representations - or "The classifier does not matter when the (text) representation is so good!". Inf. Process. Manag. 60(4) (2023) (Qualis A1; h5–index: 83, Impact Factor: 8.6)
3. A. Pereira; Felipe Viegas; M. A. Gonçalves; L. Rocha. Evaluating the Limits of the Current Evaluation Metrics for Topic Modeling. In: Webmedia 2023.. p. 119-129. (Qualis A4; h5-index: 7)
4. A. Souza Jr, P. Cecilio, **Felipe Viegas**, W. Cunha, E. Tuler, L. Rocha: Evaluating Topic Modeling Pre-processing Pipelines for Portuguese Texts. WebMedia 2022: 191-201 (Qualis A4; h5-index: 7)
5. W. Cunha, V. Mangaravite, C. Gomes, S. D. Canuto, E. Resende, C. Nascimento, **Felipe Viegas**, C. França, W. S. Martins, J. M. Almeida, T. Rosa, L. Rocha, M. A. Gonçalves: On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. Information Processing & Management 58(3): 102481 (2021) (Qualis A1; h5–index: 83, Impact Factor: 8.6)
6. W. Cunha, Sérgio D. Canuto, **Felipe Viegas**, T. Salles, C. Gomes, V. Mangaravite, E. Resende, T. Rosa, M. A. Gonçalves, L. Rocha: Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling. Information Processing & Management, 2021, 2023. 57(4): 102263 (2020) (Qualis A1; h5–index: 83, Impact Factor: 8.6)
7. W. Cunha, **Felipe Viegas**, R. Alencar, F. Mourão, T. Salles, D. Carvalho, M. A. Gonçalves, L. Rocha: A Feature-Oriented Sentiment Rating for Mobile App Reviews. ACM International World Wide Web Conference 2018: 1909-1918 (Qualis A1; h5–index: 106);

Prêmios:

- Vencedor (1o. lugar) do **Prêmio Capes de Teses 2024** - Área: Computação
- Prêmios na 8ª e 9ª Google Latin America Research Awards (Google LARA) - Taxa de aprovação 2%;
- Melhor tese de doutorado do PPGCC UFMG (2024) - Indicada ao Prêmio UFMG e ao Prêmio Capes;
- Publicações da tese (diretas e inspiradas) com mais de 200 citações (fonte: Google Scholar);
- Finalista do Concurso de Teses e Dissertações da Sociedade Brasileira de Computação (CTD SBC 2024) e do Concurso de Teses e Dissertações do Simpósio Brasileiro de Sistemas Multimídia e Web (Webmedia).

9. Agradecimentos

Essa tese foi suportada em parte por financiamento do CNPq, CAPES, FAPEMIG, FAPESP, Google e AWS.

Referências

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Griffiths, T. L., Jordan, M. I., Tenenbaum, J. B., and Blei, D. M. (2004). Hierarchical topic models and the nested chinese restaurant process. In *Advances in neural information processing systems*, pages 17–24.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Hamilton, W. L., Clark, K., Leskovec, J., and Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. *CoRR*, abs/1606.02820.

Hutto, C. J. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Li, C., Duan, Y., Wang, H., Zhang, Z., Sun, A., and Ma, Z. (2017). Enhancing topic modeling for short texts with auxiliary word embeddings. *ACM TOIS*.

Sachan, D. S., Zaheer, M., and Salakhutdinov, R. (2019). Revisiting lstm networks for semi-supervised text classification via mixed objective function. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6940–6948.

Shi, T., Kang, K., Choo, J., and Reddy, C. K. (2018). Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *WWW '18*, pages 1105–1114.

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B., and Hu, X. (2024). Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, 18(6).