Universidade de Pernambuco
Programa de Pós-Graduação em Engenharia da Computação (PPGEC)

Proposta de Tese de Doutorado

Área: Inteligência Computacional

Título: Integração de Inteligência Artificial Simbólica e LLMs para Desambiguação Lexical em Português Brasileiro com Ênfase em Regionalismos e Problemas Polissêmicos

Orientador: Profº Dr. Cleyton Mário de Oliveira Rodrigues

A desambiguação lexical de sentidos (DLS) é uma tarefa central no Processamento de Linguagem Natural (PLN), essencial para compreender o significado de palavras polissêmicas em seus contextos específicos. Embora modelos de linguagem de grande escala (LLMs) tenham avançado significativamente nessa área, pesquisas recentes evidenciam limitações importantes, como alucinações, falta de generalização em domínios específicos e ausência de explicabilidade robusta. Essas limitações tornam os LLMs insuficientes para capturar a riqueza semântica de idiomas como o português brasileiro, que inclui regionalismos, gírias e expressões idiomáticas amplamente dependentes de contexto sociocultural.

Além disso, tipos distintos de problemas polissêmicos — como palavras de sentido literal e figurado ("mão" como parte do corpo ou medida), ambiguidades morfológicas ("capital" como cidade ou recurso financeiro) e expressões contextuais ("quebrar o galho") — representam desafios adicionais para os modelos. A falta de bases de dados especializadas que reflitam essas nuances linguísticas e culturais reforça a necessidade de novas abordagens.

Este trabalho propõe uma abordagem híbrida que integra LLMs com técnicas de Inteligência Artificial Simbólica (IA Simbólica) para aprimorar a DLS no português brasileiro. A proposta inclui a criação de uma base de dados inédita e rica em expressões polissêmicas, contemplando diferentes regiões, contextos culturais e domínios específicos, como jornalismo, direito e educação. Essa base será utilizada no desenvolvimento de um modelo neuro-simbólico, que combina a capacidade contextual dos LLMs com representações simbólicas derivadas de ontologias linguísticas e regras semânticas. Essa integração visa capturar a complexidade da língua, ao mesmo tempo em que promove maior explicabilidade e precisão nas tarefas de DLS.

A metodologia será estruturada em três etapas principais: (1) construção de um corpus diversificado, anotado manualmente, destacando regionalismos, gírias e expressões idiomáticas em diferentes contextos; (2) desenvolvimento de um modelo híbrido, onde LLMs geram embeddings contextuais, complementados por regras simbólicas para validação semântica e desambiguação; e (3) avaliação comparativa do modelo proposto, utilizando métricas padrão de DLS, bem como análises qualitativas para medir sua eficácia em capturar nuances culturais e contextuais.

Os principais diferenciais incluem: (1) o desenvolvimento de uma base de dados única e especializada, cobrindo problemas polissêmicos diversos no português brasileiro; (2) a integração de técnicas simbólicas e conexionistas para melhorar explicabilidade e precisão; e (3) a validação do modelo em contextos reais, promovendo aplicações em tradução automática, assistentes virtuais, análise de sentimentos e recuperação de informações.

Espera-se que esta pesquisa contribua significativamente para o avanço da DLS em português, oferecendo uma solução robusta e explicável, capaz de capturar a riqueza semântica e cultural da língua, enquanto estabelece um novo padrão de benchmark para a área.

Referências Bibliográficas:

- **Loureiro, Daniel; Rezaee, Kiamehr; Pilehvar, Mohammad Taher; Camacho-Collados, Jose.** Analysis and Evaluation of Language Models for Word Sense Disambiguation. *Computational*

Linguistics, v. 47, n. 2, p. 387–443, 2021. Disponível em: https://aclanthology.org/2021.cl-2.14/. Acesso em: 17 jan. 2025.

- **Smythos.** Symbolic AI in Natural Language Processing: A Comprehensive Guide. *Smythos*, 2024. Disponível em: https://smythos.com/artificial-intelligence/symbolic-ai/symbolic-ai-in-natural-language-processing/. Acesso em: 17 jan. 2025.

- **Hamilton, Kyle; Nayak, Aparna; Božić, Bojan; Longo, Luca.** Is Neuro-Symbolic AI Meeting its Promise in Natural Language Processing? A Structured Review. *arXiv preprint arXiv:2202.12205*, 2022. Disponível em: https://arxiv.org/abs/2202.12205. Acesso em: 17 jan. 2025.

- **Sumanathilaka, T.G.D.K.; Micallef, Nicholas; Hough, Julian.** Can LLMs assist with Ambiguity? A Quantitative Evaluation of various Large Language Models on Word Sense Disambiguation. *arXiv preprint arXiv:2411.18337*, 2024. Disponível em: https://arxiv.org/abs/2411.18337. Acesso em: 17 jan. 2025.

- **Panchendrarajan, Rrubaa; Zubiaga, Arkaitz.** Synergizing Machine Learning & Symbolic Methods: A Survey on Hybrid Approaches to Natural Language Processing. *arXiv preprint arXiv:2401.11972*, 2024. Disponível em: https://arxiv.org/abs/2401.11972. Acesso em: 17 jan. 2025.