Extração Automática de Conhecimento da Literatura Científica Usando Grandes Modelos de Linguagem para Caracterização e Design de Materiais

Orientador: Gabriel R. Schleder
Co-orientador: Bruno Focassio

1 Introdução

Modelos de Linguagem de Grande Escala (LLMs) têm transformado rapidamente o cenário da pesquisa científica ao possibilitar novas capacidades para compreensão, geração e manipulação de linguagem natural em larga escala. Treinados em grandes volumes de texto, esses modelos capturam relações semânticas complexas e conhecimento específico de domínio, viabilizando sua aplicação em tarefas como sumarização, classificação e geração de hipóteses. Seu impacto é particularmente relevante em áreas intensivas em dados, onde processar grandes volumes de informação é essencial.

Uma parcela significativa do conhecimento científico encontra-se em formatos não estruturados, textos narrativos, tabelas e figuras, o que dificulta análises em larga escala. A curadoria manual desses dados é custosa e demorada, enquanto abordagens tradicionais baseadas em regras apresentam limitações de generalização. Nesse contexto, os LLMs oferecem uma solução promissora ao permitir a extração automatizada e a estruturação de informações relevantes diretamente da literatura científica, sem a necessidade de grandes conjuntos de dados anotados para cada domínio específico.

Este projeto propõe o desenvolvimento de um pipeline estruturado para extração de conhecimento a partir de textos científicos de ciência dos materiais, utilizando LLMs. O objetivo é transformar dados não estruturados em representações estruturadas, como registros tabulares, triplas semânticas e grafos de entidade-relação, que possam ser utilizadas em aplicações posteriores, como análise de dados experimentais, construção de grafos de conhecimento e desenvolvimento de modelos preditivos.

2 Estado da arte

A aplicação de técnicas de processamento de linguagem natural (NLP) à literatura de ciência dos materiais ganhou impulso significativo a partir de meados da década de 2010, motivada pela crescente demanda por curadoria automatizada de dados de síntese, estrutura e propriedades de materiais. O projeto MatScholar [1] foi um dos primeiros esforços sistemáticos nessa direção, desenvolvendo modelos de reconhecimento de entidades nomeadas especificamente treinados em abstracts de artigos de materiais para identificar entidades como compostos inorgânicos, propriedades, condições de aplicação e técnicas experimentais. O corpus resultante e os modelos publicados tornaram-se referências amplamente utilizadas pela comunidade.

Iniciativas subsequentes como o ChemDataExtractor [2] e o NERRE (Named Entity and Relation Recognition Extraction) [3] ampliaram o escopo da extração para incluir valores numéricos de propriedades e suas unidades, permitindo a construção automatizada de tabelas de dados experimentais a partir de texto corrido. Essas abordagens, embora eficazes em contextos específicos, dependiam de heurística para extração automatizada de padrões via expressões regulares ou de modelos supervisionados com anotação extensiva e generalizavam mal para subdomínios não contemplados no treinamento, como novos sistemas de ligas ou classes de materiais funcionais emergentes.

O advento de modelos de linguagem baseados em Transformers abriu novas possibilidades para o domínio. O MatBERT [4] e o MatSciBERT [5] foram pré-treinados em grandes coleções de artigos de ciência dos materiais, demonstrando ganhos expressivos em tarefas de NER (Named Entity

Recognition) e classificação de propriedades em relação a modelos de domínio geral como BERT e SciBERT [6, 7]. Esses modelos estabeleceram que o pré-treinamento em literatura especializada é determinante para o desempenho em extração de informações técnicas altamente específicas.

Mais recentemente, LLMs como GPT-4 e Claude têm sido avaliados em tarefas de extração em ciência dos materiais em regime zero-shot e few-shot. Estudos como o de Zheng et al. [8] e Dagdelen et al. [9] demonstraram que esses modelos conseguem extrair com boa precisão condições de síntese e propriedades de materiais a partir de parágrafos experimentais, sem treinamento adicional, desde que os prompts sejam cuidadosamente elaborados. Contudo, os autores também identificaram limitações relevantes em precisão numérica e na extração de relações quantitativas complexas, apontando para a necessidade de estratégias de validação pós-extração.

Apesar desses avanços, desafios centrais permanecem: (i) a extração confiável de valores numéricos com suas unidades e incertezas, (ii) a resolução de ambiguidades na nomenclatura de compostos, (iii) a rastreabilidade das informações extraídas até as passagens originais dos artigos, (iv) o custo e escalabilidade de modelos consolidados em plataformas hospedadas em nuvem contra modelos hospedados localmente, (v) e a precisão e generalização das informações extraídas no corpus coletado diante de todo o domínio do conhecimento. Este projeto busca contribuir para esses desafios por meio do desenvolvimento e avaliação de um pipeline baseado em LLMs modernos, com foco na extração estruturada de propriedades e condições experimentais a partir da literatura de ciência dos materiais.

3 Objetivos

O objetivo geral do projeto é desenvolver e avaliar um pipeline estruturado para extração automatizada de informações a partir da literatura científica utilizando modelos de linguagem de grande escala. Dentre os objetivos específicos temos:

- Projetar um sistema de classificação para identificar documentos ou trechos relevantes;
- Implementar um módulo de extração baseado em LLMs para conversão de texto não estruturado em dados estruturados;
- Avaliar a acurácia e consistência do processo de extração;
- Construir um conjunto de dados estruturados como prova de conceito;
- Documentar o pipeline visando reprodutibilidade e extensões futuras.

4 Metodologia

O projeto será desenvolvido seguindo uma abordagem modular inspirada em Knowledge Extraction Pipelines (KEP) [10], com foco nas etapas de classificação e extração, utilizando principalmente LLMs pequenos.

4.1 Coleta de Dados: Seleção de um corpus de artigos científicos (resumos ou textos completos) e aplicação de etapas de pré-processamento, incluindo limpeza e segmentação textual.

4.2 Etapa de Classificação: Utilização de LLMs para classificar documentos ou trechos de texto quanto à relevância para o problema de interesse. Serão exploradas estratégias baseadas em prompts e extração baseada em padrões.

4.3 Etapa de Extração: Definição de templates estruturados contendo as variáveis de interesse. Os LLMs serão utilizados para extrair informações estruturadas a partir dos textos selecionados, com uso de técnicas de engenharia de prompts para melhorar a consistência dos resultados.

4.4 Pós-processamento e Validação: Normalização dos dados extraídos (unidades, formatos e terminologias), bem como aplicação de verificações de consistência. Quando possível, os resultados serão comparados com um subconjunto anotado manualmente.

4.5 Armazenamento e Recuperação: Armazenamento dos dados estruturados em banco de dados ou sistema de busca vetorial, permitindo consultas e uso em aplicações posteriores.
4.6 Avaliação: Definição de métricas de desempenho, incluindo precisão, recall, completude estrutural e consistência entre execuções. Será realizada análise quantitativa e qualitativa dos resultados.

5 Cronograma

Atividades	Meses											
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
Revisão bibliográfica												
Coleta e pré-processamento												
Módulo de classificação												
Módulo de extração												
Integração do pipeline												
Validação e avaliação												
Dataset e documentação												
Relatório final												

Referências

[1] L. Weston, V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, K. A. Persson, G. Ceder e A. Jain, “Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature”, J. Chem. Inf. Model. **59**, 3692 (2019).

[2] M. C. Swain e J. M. Cole, “ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature”, J. Chem. Inf. Model. **56**, 1894 (2016).

[3] Z. Nasar, S. W. Jaffry e M. K. Malik, “Named Entity Recognition and Relation Extraction: State-of-the-Art”, ACM Comput. Surv. **54** (2021).

[4] A. Trewartha, N. Walker, H. Huo, S. Lee, K. Cruse, J. Dagdelen, A. Dunn, K. A. Persson, G. Ceder e A. Jain, “Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science”, Patterns **3**, 100488 (2022).

[5] T. Gupta, M. Zaki, N. M. A. Krishnan e Mausam, “MatSciBERT: A materials domain language model for text mining and information extraction”, npj Comput. Mater. **8**, 102 (2022).

[6] J. Devlin, M.-W. Chang, K. Lee e K. Toutanova, “Bert: pre-training of deep bidirectional transformers for language understanding”, arXiv:1810.04805 (2019).

[7] I. Beltagy, K. Lo e A. Cohan, “Scibert: a pretrained language model for scientific text”, arXiv:1903.10676 (2019).

[8] C. Zheng, H. Zhou, F. Meng, J. Zhou e M. Huang, “Large Language Models Are Not Robust Multiple Choice Selectors”, arXiv 2309.03882 (2024).

[9] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson e A. Jain, “Structured information extraction from scientific text with large language models”, Nature Comm. **15**, 1418 (2024).

[10] V. T. da Silva, A. Rademaker, K. Lionti, R. Giro, G. Lima, S. Fiorini, M. Archanjo, B. W. Carvalho, R. Neumann, A. Souza, J. P. Souza, G. de Valnisio, C. N. Paz, R. Cerqueira e M. Steiner, “Automated, llm enabled extraction of synthesis details for reticular materials from scientific literature”, arXiv:2411.03484 (2024).