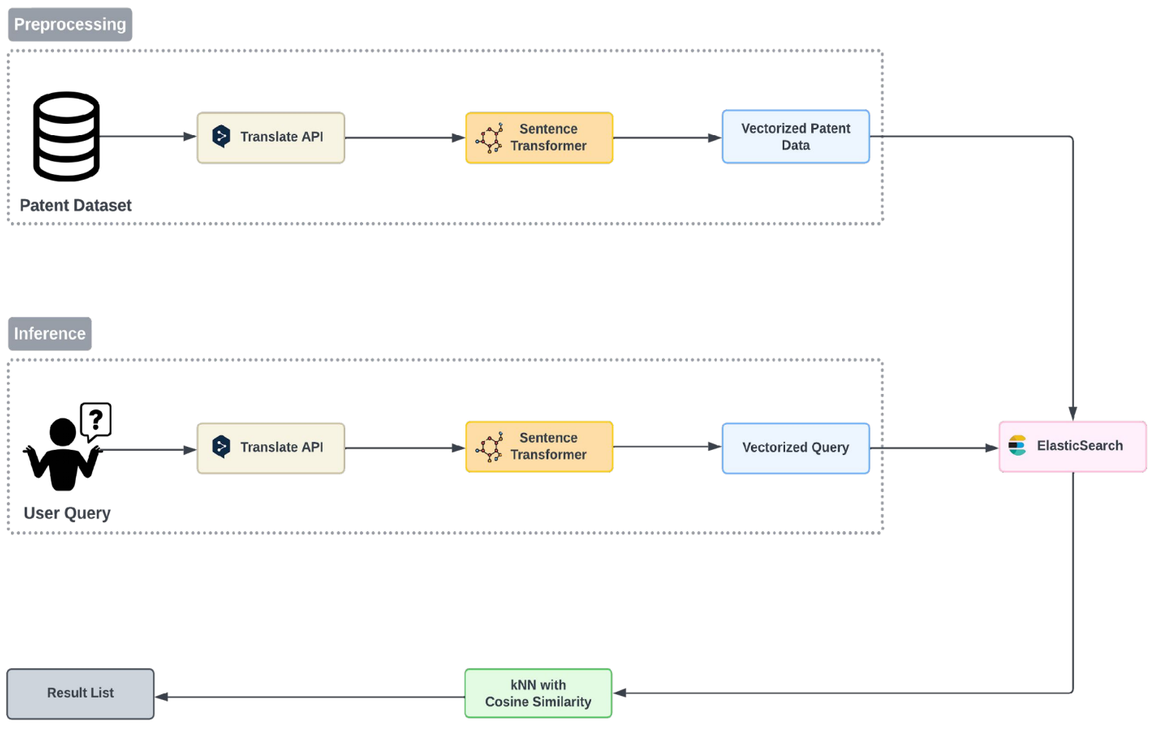
Artificial Intelligence Based Patent Search System

Buğra Şimşek 2220356187
Ali Kemal Tamkoç 2200356856
Onur Demirel 2220356189
Behiç Kılınçkaya 2220356192

Main findings

- Classical information retrieval methods, such as TF-IDF, **can't capture** the nuances of patent data due to the complexity of the language used in the domain.
- Language models leveraging **transformer architecture and attention mechanisms have demonstrated high performance** in the task of patent search.
- **Domain-specific training** of language models significantly improves their performance.



Introduction

The ever-expanding landscape of patents presents a significant challenge for researchers and innovators. Patent offices around the world report record-breaking numbers of patent applications every year[1,2], including Turkey[3]. This surge makes efficiently navigating this vast and crucial resource paramount. Our system addresses this challenge by leveraging modern natural language processing (NLP) and information retrieval techniques.

Methods

The system utilizes a custom dataset of aerospace-related patent documents obtained from the USPTO database. This dataset focuses on patents filed by the top 100 aerospace companies[7], representing a significant portion of the industry. The text data undergoes pre-processing steps including translation to English, removing patents with missing data sections, cleaning irrelevant characters and figures. Following pre-processing, a transformer model specifically trained on patent data, PatentSBERTa, is employed to convert the text into vector representations of dimensionality 768. These vector representations capture semantic meaning with the help of attention mechanisms, enabling the system to understand the relationships between concepts within the patents. ElasticSearch[8] acts as a vector database and facilitates efficient vector-based search. To retrieve the most relevant patents, the system utilizes a KNN (k-Nearest Neighbors) algorithm with cosine similarity.

Results

Our system with domain-specific model, fine-tuned on patents, achieved superior performance compared to general-purpose language models and classical methods such as TF-IDF[4] (in terms of precision, recall, MRR). PatentSBERTa excelled at:

- retrieving relevant patents for real-world queries (precision: 67.4% - recall: 46.6%),
- synthetic queries created by LLMs (precision: 61.7% - recall: 63.5%)
- and patent research reports prepared by experts (precision: 41.6% - recall: 40.2%).

This success is attributed to its ability to capture the nuances of patent language through domain-specific fine-tuning.

Approach	Precision	Recall	MRR
Expert Queries			
TF-IDF	33.50%	9.70%	0.25
MiniLM	55.70%	39.50%	0.635
PatentSBERTa	67.40%	46.60%	0.802
Research Reports			
TF-IDF	17.90%	12.10%	0.098
MiniLM	29.30%	32.80%	0.273
PatentSBERTa	41.60%	40.20%	0.359
Language Model Queries			
TF-IDF	25.90%	22.20%	0.413
MiniLM	56.40%	49.30%	0.637
PatentSBERTa	61.70%	63.50%	0.769

Discussion

This study demonstrates the effectiveness of our AI-driven patent search system utilizing PatentSBERTa. The system outperforms traditional methods, but limitations exist, such as focusing only on aerospace patents and lack of a large labeled dataset. Future improvements include expanding to more technical domains, integrating the search system with a patent classification model, adding multimodal capabilities, and investigating user-driven query reformulation. Overall, it offers a promising solution for efficient and accurate patent search, particularly in specialized fields, but is open to improvement.

References

1. Office of Chief Economist, "Intellectual property and the U.S. economy: Third edition," www.uspto.gov, Mar. 17, 2022. https://www.uspto.gov/ip-policy/economic-research/intellectual-property-and-us-economy (accessed Feb. 10, 2024)
2. The World Intellectual Property Organization (WIPO), "World Intellectual Property Indicators Report: Record Number of Patent Applications Filed Worldwide in 2022," The World Intellectual Property Organization (WIPO), Geneva, Nov. 2023. Accessed: Feb. 10, 2024. [Online]. Available: https://www.wipo.int/pressroom/en/articles/2023/article_0013.html
3. Türk Patent ve Marka Kurumu, "2022 Yılı Sınai Mülkiyet Sayıları Açıklandı," www.turkpatent.gov.tr, Feb. 01, 2023. [Online]. Available: https://www.turkpatent.gov.tr/haberler/2022-yili-sinai-mulkiyet-sayilari-aciklandi (accessed Feb. 10, 2024).
4. G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing & Management, vol. 24, no. 5, pp. 513–523, Jan. 1988, doi: https://doi.org/10.1016/0306-4573(88)90021-0.
5. A. Souili, D. Cavallucci, and F. Rousselot, "Natural Language Processing (NLP) – A Solution for Knowledge Extraction from Patent Unstructured Data," Procedia Engineering, vol. 131, pp. 635–643, Dec. 2015, doi: https://doi.org/10.1016/j.proeng.2015.12.457.
6. S. Marcos-Pablos and F. J. García-Peñalvo, "Information retrieval methodology for aiding scientific database search," Soft Computing, vol. 24, no. 8, pp. 5551–5560, Oct. 2018, doi: https://doi.org/10.1007/s00500-018-3568-0.
7. M. Morrison, "Top 100 aerospace companies ranked by revenue," *Flight Global*. https://www.flightglobal.com/flight-international/top-100-aerospace-companies-ranked-by-revenue/154606.article (accessed Feb. 06, 2024)
8. Elastic, "Elasticsearch Guide [8.2]," www.elastic.co. https://www.elastic.co/guide/en/elasticsearch/reference/current/index.html (accessed Feb. 16, 2024).