PASTA: Strategic Recipe Guiding the Next PhD Years

Renukswamy Chikkamath Hochschule München
Supervised by: Markus Endres

Abstract: This paper outlines the motivation and necessity for tackling challenges in prior art search. Drawing from our DFG proposal, we detail planned milestones for a two-year development of an AI-based prototype for prior art search. The prototype, aligned with the Ph.D. thesis, integrates a timeline and architectural flow. The envisioned prototype covers research and development at crucial stages in AI-based search, the prototype focuses on optimal patent representations, efficient document retrieval, and semantic re-ranking. A novelty model based on examination reports is explored for re-ranking, enhanced by semantic filters. A substantive examination model is proposed for in-page semantic search, reducing human effort by cross-questioning and highlighting relevant paragraphs. The paper concludes with considerations for future enhancements, ensuring an adaptable tool in the evolving landscape of AI-based prior art search.

Keywords: Artificial Intelligence, Prior Art Search, Patent Information Retrieval, Semantic Search.

1 Introduction

The general patent life cycle begins with an innovative idea, prompting patent applicants to conduct thorough patentability searches to confirm its novelty. Attorneys then assist in drafting and submitting a patent application to the patent office. After a preliminary check, the application undergoes detailed examination, including a prior art search, to verify novelty and non-obviousness, and it becomes public. If the drafted idea is deemed novel and non-obvious, a patent is granted; otherwise, the application is rejected. Third-party observations outside the patent office may lead to appeals, involving litigation searches or analysis. In summary, patent searches in both patent and non-patent literature are crucial in the process of obtaining a patent for an idea. Even a century-old painting on a wall describing an idea can negate novelty and invalidate a patent, making the search for prior art akin to finding a needle in a haystack without AI or ML.

The process of obtaining a patent involves various stages, including application and examination, which encompass patent analysis tasks like classification and prior art search. However, the exponential growth in patent document volume has made manual analysis time-consuming and labor-intensive. Keyword-based approaches, commonly used for semi-automated analysis, yield sub-optimal results due to the overwhelming volume of patents. Assessing the novelty and non-obviousness of the claimed invention in a patent application requires the identification of relevant prior art – in other words, state-of-the-art or background art. Prior art is any available evidence presented in readable formats, potentially including simplified figures, predating the application submission. Prior art search, a crucial step conducted before and after patent application, has become challenging with these inadequate methods targeting large patent volumes. The extensive volume of patent applications demands significant human labor and domain expertise for analysis and prior art search.

Even though there are traces in the state of the art (refer to Section 1.1 of the DFG proposal) in using AI for prior art search, optimal vector representation of patents and memory footprint issues remain unanswered questions for efficient prior art search as claimed by European Patent Office (EPO)[1]. Even when AI[1] is applied[2] exclusively for semantic search, it typically employs only the abstracts or claims of patents to create vector representations. This approach leads to compromised precision and recall, making it suboptimal. Apart from representation challenges, these initiatives lack a substantive examination tool that utilizes explainable AI. Furthermore, the emphasis of these initiatives is mostly on syntactic filters or metadata-based filters during the search process. The demand for semantic filters is much greater, incorporating entities and their relationships within the text, as well as novelty metrics to generate the most relevant initial search outcomes.

Similarly, there has been significant development in automating prior art search using AI, exemplified by the AT&T initiative known as the PQAI project[3]. Given that PQAI is open source, we foster close collaboration by sharing common interests, such as investigating the

[1] https://www.linkedin.com/pulse/ai-guided-cpc-classification-alexander-klenner-bajaja/
[2] https://www.wipo.int/about-ip/en/artificial_intelligence/search.jsp
[3] https://search.projectpq.ai/

well-formedness of search queries in DL-based search engines. While the PQAI project aims to assist prior art searchers, it still faces core challenges, particularly in the realm of patent representation. Our project proposal distinguishes itself in several ways (refer to pg. 3 of DFG proposal).

In the era of ChatGPT[4], its capabilities and potential in patent semantic search are limited. Once, OpenAI CEO Sam Altman said[5], "ChatGPT is cool but a horrible product", implying that the integration is not well-designed yet. Moreover, empirical evidence suggests that ChatGPT is not suitable for direct utilization in prior art searches. Patent professionals express concerns that models like ChatGPT may resemble "stochastic parrots" for examination purposes, lacking critical thinking abilities necessary for patent prior art search. In contrast, we propose to investigate how, why, or whether ChatGPT-like Large Language models (LLMs) can be used for prior art search. In particular, we aim to investigate state-of-the-art LLMs to address the novelty identification task given sets of pairs of patent text.

Unlikely as it may seem, our proposed approach revolves around the development of optimal representation vectors for patents, utilizing semantic filters, and providing tools for substantive examination through explainable AI. In the subsequent section, we added the objectives inherent to our proposed approach.

2 Objectives

In light of the state-of-the-art transforming an overlooked search into a comprehensive one in a timely manner requires improvement in several areas that align with our objectives. This includes:

1. Enhancing the way machines/models understand patents through representation learning, where the optimal machine-readable representation of patent text is discovered (referred to as the representation module). Here we develop an embedding model and a retriever.
2. Providing qualitative feedback to users as relevance signals by implementing semantic filters to enhance the precision of prior art search (referred to as the semantic filters module). In this module, we develop DL models to identify novelty.
3. Facilitating substantive examination of patents by improving the interpretability of search results with greater flexibility to engage users in steering directed and explainable prior art search (referred to as the examination module). In this module, we develop a DL model for in-page semantic search, and automatically highlight technical aspects within the patent document.
4. Developing an AI-based prior art search tool (refer to Figure 1) focused on patent information retrieval at various levels, including sentence, passage, and document (referred to as the interactive end-user application accommodating previously mentioned modules). In this application, we integrate the previously mentioned DL models into the search tool.

[4] https://openai.com/blog/chatgpt

[5] https://futurism.com/the-byte/ceo-openai-chatgpt-horrible-product

3 Preliminary Work

The preliminary work for this proposal is characterized by a unique approach and a focus on critical research areas, contributing significantly to the foundation of the PhD thesis. Our experimental case study on patent novelty detection [2], using machine learning (ML) and deep learning (DL) algorithms with data from EPO search reports [3], revealed the innovative finding that classical ML models demonstrated comparable accuracy and robustness to DL models. Recognizing the potential for dataset expansion to automate manual novelty identification, insights from this investigation will inform the development of substantive examination models, with a long-term objective of creating a universal novelty detection model.

Additionally, a comprehensive survey article on DL techniques [4] summarized state-of-the-art approaches, guiding all milestones (M) in this project. To streamline semantic annotation in manual patent analysis, we proposed a novel dataset and ML algorithms for automated highlighting [5]. This preliminary work included creating a dataset with 150k samples from USPTO patents, conducting exploratory data analysis, and developing ML models dedicated to highlighting technical aspects within patent paragraphs. In addition, we demonstrated the significance of domain-specific language model through the application of patent classification. Our approach surpassed the state-of-the-art methods at the time of publication, and we contributed a new dataset for classification purposes [6]. To facilitate testing by domain experts, we developed browser extensions for in-page semantic search and highlighting of patents in Google Patents. These extensions enhance the exploration and analysis of patent documents within the browser. The semantic annotations ([5, 7]) are particularly relevant for substantive examination. Further, an investigation into natural language search query parameters, grammar, verbosity, and specificity [8], as well as an exploration of topic models for information retrieval [9], were conducted to understand DL-based search

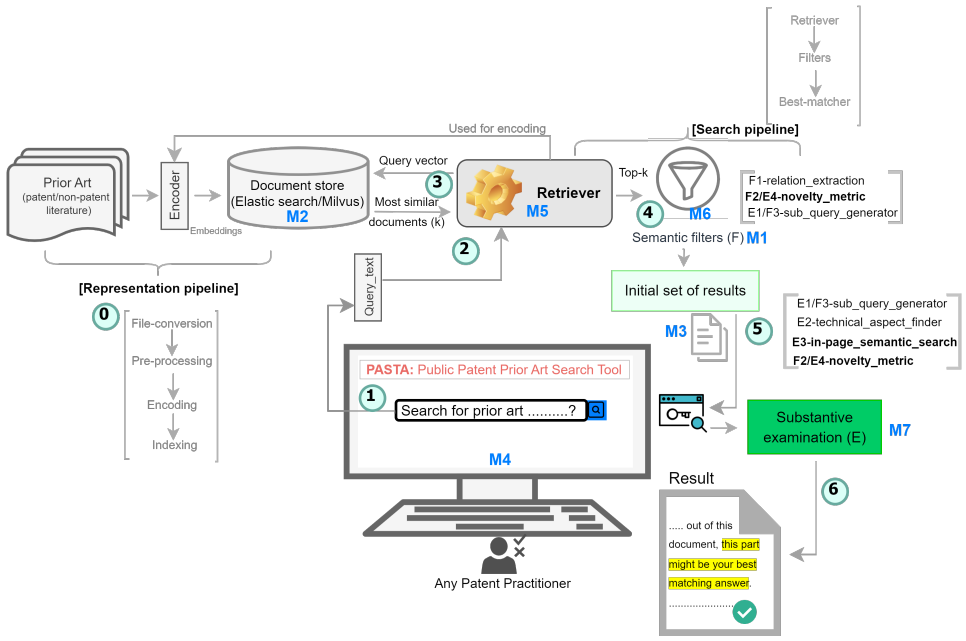

Figure 1: *PASTA architecture showing workflow of AI based prior art search prototype*

engine sensitivity, with implications for all milestones proposed.

A detailed architecture flow showing the components of the search prototype and a research timeline for 2 years are given in the next Section.

4 Research Plan

This section provides a detailed architecture along with an annotated timeline chart for prioritizing the milestones of the Ph.D. thesis. At the abstract level depicted in Figure 1, a general search process begins with a given query from the user, as shown in the center of the figure (step 1; steps are numbered within the circle). Before this, in the preparatory step (step 0), our architecture is primarily founded on an embedding retrieval method. This method involves an embedding model trained on optimal representations of patent text in a sentence-transformer fashion, serving as both a retriever and an encoder. Documents to be searched are encoded using this encoder and stored in a vector document store.

The query text is then passed to the retriever in step 2. The retriever performs the search in the document store to retrieve the most similar documents in step 3. Once the top-k relevant documents are retrieved by the retriever, they are not directly shown to the user, as in prevailing state-of-the-art approaches. Instead, they are fed into semantic search filters, allowing the user to assess their quality in step 4. In step 4, the user can apply several filters, such as the novelty metric, to restrict results based on the novelty of the search query. Further, in step 5, examination modules help improve readability and explainability based on in-page semantic search. Finally, in step 6, the user can be given a final result highlighting the answer or text relevant to the search query. The architecture also contains milestones (M), which are discussed in the next section.

5 Upcoming Milestones

This section details the milestones covered in the consecutive years. The detailed timeline in reference to these milestones is described in Figure 2.

5.1 M1 - Exploring LLMs for Novelty Detection (ChatGPT Vs. GoogleBARD)

In this milestone, our objective is to construct a small test set utilizing patent examination reports. The aim is to assess the capabilities and challenges faced by Large Language Models (LLMs), specifically ChatGPT and GoogleBARD, in discerning novelty between a patent and given prior art. Additionally, we conduct a comparative analysis between the effectiveness of LLMs and state-of-the-art embedding models using similarity metrics. This task provides valuable insights into the development of representation vectors and patent novelty credibility for patent prior art search.

5.2 M2 - Investigate and Develop Benchmark Dataset for Patent Information Retrieval

In the current state of the art, there is a shortage of well-designed patent information retrieval datasets suitable for testing AI-based search engines. Therefore, we aim to create a mid-scale patent retrieval dataset

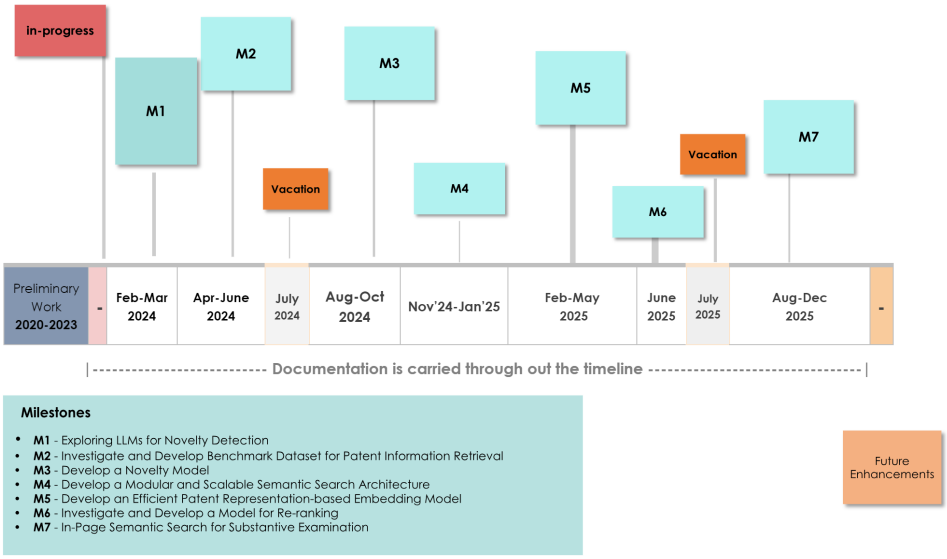

Figure 2: *PhD thesis timeline including milestones annotated*

that can be universally utilized to evaluate any AI-based search tools. This dataset will be employed initially for testing on our search prototype.

5.3 M3 - Develop a Novelty Model

Building upon the knowledge acquired in M1, this milestone focuses on crafting a novelty detection model. This model will serve as a semantic filter for our search tool. Additionally, we aim to investigate the integration of novelty-based models in re-ranking the retrieved results based on their novelty relevance. Hence, M3 is intricately linked with the research and development aspects of M1 and M6.

5.4 M4 - Develop a Modular and Scalable Semantic Search Architecture

In this milestone, our objective is to construct an AI-based semantic search architecture. The architecture is bifurcated into two primary components: the representation pipeline and the search pipeline. We will also develop a basic front-end interface for query input and result display using front-end technologies. This modular design facilitates the experimentation and validation of various combinations of preprocessing, representation, filtering, and examination models. Over time, our modular architecture allows for continuous upgrades to accommodate evolving technologies and optimal models. This milestone enables seamless communication between the tool and all other models for end-to-end prior art search.

5.5 M5 - Develop an Efficient Patent Representation-based Embedding Model

In this milestone, our objective is to identify and develop an embedding model for representing patent text in vector form. Given the substantial length of patents, determining which segments are crucial for conversion into vectors for semantic search poses a challenge. To address this, we initially test M4 using state-of-the-art general-purpose embedding models on various parts of patents, such as abstracts, claims, and combinations of other relevant paragraphs, to represent the details of the invention in one or more vectors. This preliminary exploration provides insights into which segments effectively capture the essence of an invention. Subsequently, by selecting the most efficient parts of patents and leveraging labeled examination search reports from examiners, we aim to train or fine-tune the model to create an efficient embedding model. This embedding model will be employed to process raw patents and store them in vector databases for search purposes.

5.6 M6 - Investigate and Develop a Model for Re-ranking

In this milestone, our focus is on investigating and developing the optimal model for re-ranking search results based on a given query. The initial set of relevant results obtained from the retriever, based on embedding relevance, is input to the re-ranker. The re-ranker then selects the most relevant subset from this initial set. M6 collaborates closely with M3, exploring how novelty models behave in serving as effective re-rankers, mimicking an examiner's selection of the closest documents.

5.7 M7 - In-Page Semantic Search for Substantive Examination

Despite obtaining a handful of results from M6, the arduous task of individually reading and verifying these results by patent searchers remains. To address this challenge, we aim to develop an in-page semantic search model. This involves creating a patent question-answering dataset where, given a context, a question needs to be answered, and relevant content, paragraphs, phrases, or text segments within that context are highlighted. A model trained on such a dataset will be proficient in in-page semantic search, empowering patent searchers to cross-question prior art documents.

An illustrative usage scenario of the in-page search facilitating substantive examination is as follows: Imagine a patent application is identified in the relevant prior art set after M6. When a user is interested in exploring and verifying its relevance to the query or finding the technical aspects crucial for determining the inventiveness of the searched query within this prior art, the user can click on "Substantive Examination." This action opens a new browser tab displaying the prior art in Google Patents. Implicitly, a browser extension is activated for in-page semantic search, utilizing the initially used text for the search. Relevant paragraphs or technical aspects are highlighted, as shown in Figure 21 above. The user is now free to explore with more cross-questions to understand or verify the contents of the patent. Unlike keyword-based search, context and semantic-based answers are extracted from prior art, even if the content is drafted using different terminologies common in the patent domain.

6 Conclusion and Future Enhancements

This paper provides a concise exploration of the motivation driving the adoption of AI in addressing challenges inherent in traditional prior art search systems. We elucidate our efforts directed towards overcoming these challenges, encompassing both preliminary work and proposed milestones. To enhance comprehension of our objectives, we present a detailed architecture, accompanied by a timeline chart illustrating the anticipated progress of milestones over the next two years.

Looking ahead, several avenues for future enhancements (F1, E1/F3, and E2 of Figure 1) to our search tool are conceivable. One potential enhancement involves optimizing the document retrieval process depicted in Figure 1. After reaching Step 3, instead of comparing millions of documents in the document store, the retriever can incorporate intelligent topic models to streamline the search and focus on a pertinent bundle of documents. Additionally, in Step 4, users can apply various filters, such as 'generate subqueries,' and refine results based on relation extraction related to key elements of the search query.

Furthermore, in Step 5, examination modules are proposed to enhance readability through a technical aspects highlighter and improve search result navigability using a sub-query generator based on combinatorial search. These supplementary models aim to instill user confidence in a specific search result by validating content used for novelty detection. Therefore, the incorporation of novelty metrics and a sub-query generator in both Step 4 and Step 5 stands out as a promising avenue for future enhancements in this research.

References

[1] Konrad Vowinckel and Volker D Hähnke. "SEARCHFORMER: Semantic patent embeddings by siamese transformers for prior art search". In: *World Patent Information* 73 (2023), p. 102192.

[2] Renukswamy Chikkamath et al. "An empirical study on patent novelty detection: A novel approach using machine learning and natural language processing". In: *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE. 2020, pp. 1–7.

[3] Julian Risch et al. "Patentmatch: a dataset for matching patent claims & prior art". In: *arXiv preprint arXiv:2012.13919* (2020).

[4] Ralf Krestel et al. "A survey on deep learning for patent analysis". In: *World Patent Information* 65 (2021), p. 102035.

[5] Renukswamy Chikkamath et al. "Patent Sentiment Analysis to Highlight Patent Paragraphs". In: *arXiv preprint arXiv:2111.09741* (2021).

[6] Renukswamy Chikkamath et al. "Patent Classification Using BERT-for-Patents on USPTO". In: *Proceedings of the 2022 5th International Conference on Machine Learning and Natural Language Processing*. 2022, pp. 20–28.

[7] Renukswamy Chikkamath et al. "Explainable Artificial Intelligence for Highlighting and Searching in Patent Text". In: *PatentSemTech2́3: 4th Workshop on Patent Text Mining and Semantic Technologies, colocated with the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 27th, 2023, Taipei, Taiwan.* (2023).

[8] Renukswamy Chikkamath et al. “Is your search query well-formed? A natural query understanding for patent prior art search”. In: *World Patent Information* 76 (2024), p. 102254.

[9] Renukswamy Chikkamath and Markus Endres. “Decoding Health Informatics Patents: Investigating Topic Models for Patent Information Retrieval”. In: *submitted to HINT́24* (2024).