PatentMatch: A Dataset for Matching Patent Claims & Prior Art

Julian Risch
Hasso Plattner Institute
University of Potsdam, Germany
julian.risch@hpi.de

Nicolas Alder
Hasso Plattner Institute
University of Potsdam, Germany
nicolas.alder@student.hpi.de

Christoph Hewel
BETTEN & RESCH Patent- und Rechtsanwälte PartGmbB
Munich, Germany
c.hewel@bettenpat.com

Ralf Krestel
Hasso Plattner Institute
University of Potsdam, Germany
ralf.krestel@hpi.de

ABSTRACT

Patent examiners need to solve a complex information retrieval task when they assess the novelty and inventive step of claims made in a patent application. Given a claim, they search for *prior art*, which comprises all relevant publicly available information. This time-consuming task requires a deep understanding of the respective technical domain and the patent-domain-specific language. For these reasons, we address the computer-assisted search for prior art by creating a training dataset for supervised machine learning called PatentMatch. It contains pairs of claims from patent applications and semantically corresponding text passages of different degrees from cited patent documents. Each pair has been labeled by technically-skilled patent examiners from the European Patent Office. Accordingly, the label indicates the degree of semantic correspondence (matching), i.e., whether the text passage is prejudicial to the novelty of the claimed invention or not. Preliminary experiments using a baseline system show that PatentMatch can indeed be used for training a binary text pair classifier and a dense passage retriever on this challenging information retrieval task. The dataset is available online: https://hpi.de/naumann/s/patentmatch.

CCS CONCEPTS

• **Computing methodologies** → Language resources; *Supervised learning*; • **Social and professional topics** → **Patents**; • **Information systems** → *Retrieval tasks and goals*.

KEYWORDS

patent documents, document classification, dataset, prior art search, dense passage retrieval, deep learning

1 PASSAGE RETRIEVAL FROM PRIOR ART

Language understanding is a very difficult task. Even more so when considering technical, patent-domain-specific documents. Modern deep learning approaches come close in grasping the semantic meaning of simple texts, but require a huge amount of training data. We provide a large annotated dataset of patent claims and corresponding prior art, which not only can be used to train machine learning algorithms to recommend suitable passages to human experts, but also illustrates how experts solve this very complex IR-problem.

In general, a patent entitles the patent owner to exclude others from making, using, or selling an invention. For this purpose, the patent comprises so-called patent claims (usually at the end of a technical description of the invention). These claims legally specify the scope of protection of the invention.To be even more precise, the legally relevant definition can be found in the independent claims, i.e., usually in claim No. 1. Said claim 1 may be only a few lines long and may comprise only rather generalized terms, in order to keep the scope of protection as broad as possible. There may be more than one independent claim, e.g., an independent system claim 1 and an independent method claim 15. The further claims are so-called dependent claims, i.e., they depend on an independent claim. This dependency is explicitly defined in the preamble of the dependent claim, e.g. by starting with: "2. The system according to claim 1, wherein…". The function of dependent claims is to define optional features of the invention, which are preferable but not mandatory for the invention (e.g., "…wherein the light source is an OLED").

In order to obtain a patent, it is required that the invention as defined in the claims is new and inventive over prior art [19]. A patent application therefore has to be filed at a patent office where it is examined on *novelty* and *inventive step* by a technically skilled examiner. In case a patent is granted, said patent is published again as a separate patent document. For this reason, there exists a huge corpus of publicly available patent documents, i.e., published patent applications and patents.

As a further consequence of this huge patent literature corpus, the examiners usually focus their prior art search on relevant patent documents. Accordingly, they try to retrieve at least one older patent document that discloses the complete invention as defined in the claims, in particular in independent claim 1. In other words, such a novelty-destroying document must comprise passages that semantically match with the definition of claim 1 of the examined patent application. Said novelty-destroying document is manually marked by an expert as *"X"* document in the search report issued by the patent office [17]. Any retrieved document that does not disclose the complete invention defined in claim 1 but at least renders it obvious, is marked as *"Y"* document in the search report. Further found documents that form technological background but are not relevant to the novelty or inventive step of claim 1, are marked as

"A" documents. As a consequence, only one retrieved *"X"* document or *"Y"* document is enough to refuse claim 1 and hence the patent application. Due to this circumstance, the search task is rather focused on precision than on recall. Usually, a search report issued for an examined patent application only comprises a few (e.g., 5) cited patent documents, wherein (as far as possible) at least one document is novelty destroying (marked as *"X"* document).

Advantageously, a search report issued by the European Patent Office (EPO) not only cites patent documents deemed relevant by an expert but also indicates for each cited document which paragraphs within the document are found to be relevant for the examined claims. Figure 1 exemplifies such a search report. The EPO search report annotates each claim of the examined patent application with specific text passages (i.e., paragraphs) of a cited document. The EPO calls this rich-format citation. Given the application with the filing number *EP18214053*, a patent officer cited prior art with the publication number *EP1351172A1*. For example, paragraphs 27-28, 60 and 70-74 are relevant passages for assessing the novelty of claims 1 and 3 to 9 (marked by an *"X"*). Furthermore, said paragraphs are also relevant for the inventive step of claim 2 (marked by an *"Y"*). The search report also lists which search terms were used. In this case, it is the IPC subclass *G06K*.

2 RELATED WORK

Finding relevant prior art is even for well-trained experts a hard and cumbersome task [10]. Due to the large volume of literature to be considered as well as the required domain knowledge, patent officers rely on modern information systems to support them with their task [18]. Nevertheless, the outcome of a prior art search, either to check for patentability or validity of a patent, remains imperfect and biased based on the patent examiner and her search strategy [15]. In addition, different patent offices can reach different conclusions for the same search [19]. With this paper we hope to open the door to qualitatively and systematically analyse the search practice particularly at European Patent Office.

Traditionally, related work at the intersection of information retrieval and patent analysis aims to support the experts by automatically identifying technical terms in patent documents [11] or keywords that relate to the novelty of claims in applications [24]. A challenge that all natural language processing applications in the patent domain have is to cope with the legal jargon and specialized terminology, which led to the use of patent-domain-specific word embeddings in deep learning approaches [1, 22]. Further, patent classification is the most prominent task for the application of natural language processing in this domain, with supervised deep learning approaches outperforming all other methods [16, 22]. Large amounts of labeled training data are available for this task because every published patent document and application is classified according to standardized, hierarchical classification schemes.

Prior art search is a document retrieval task where the goal is to find related work for a given patent document or application. Formulating the corresponding search query is a research challenge typically addressed with keyword extraction [8, 25, 27]. Further, there is research on tools to support expert users in defining search queries [23] or non-expert users in exploring the search space step

by step [14]. The task that we focus on in this paper is patent passage retrieval. Given a query passage, e.g., a claim, the task is to find relevant passages in a corpus of text documents to, e.g., decide on the novelty of the claim. In the CLEF-IP series of shared tasks, there was a *claims to passage task* in 2012 [7, 21]. The shared task dataset contains 2.3 million documents and 2700 relevance judgements of passages for training, which were manually extracted from search reports. The passages are contained in *"X"* documents and *"Y"* documents referenced by patent examiners in the search reports. Similar passage retrieval tasks can be found in other domains as well, e.g., passage retrieval for question answering within Wikipedia [3]. To the best of our knowledge, the dense passage retrieval (DPR) model for open-domain question answering by Karpukhin et al. [12] has not been used in the patent domain so far and we are the first to train a DPR model on patent data, which we describe in one of our preliminary experiments. Research in the patent domain is limited for three reasons: patent-domain-specific knowledge is necessary to understand (1) different types of documents (patent applications, granted patents, search reports), (2) different classification schemes (IPC, CPC, USPC) and (3) the steps of the patenting process (filing, examination, publication, granting, opposition).

In this paper, we present PatentMatch, a dataset of claims from patent applications matched with paragraphs from prior art, e.g., published patent documents. Professional patent examiners labeled the claims with references to paragraphs that are prejudicial to the novelty of the claim (*"X"* documents, positive samples) or that are not prejudicial but represent merely technical background (*"A"* documents, negative samples). We collected these labels from search reports created by patent examiners, resolved the claims and paragraphs referenced therein, and extracted the corresponding text passages from the patent documents. This procedure resulted in a dataset of six million examined claims and semantically corresponding (matching) text passages that are prejudicial or not prejudicial to the novelty of the claims. The remainder of this paper is structured as follows: Section 3 describes the data collection and processing steps in detail and provides dataset examples and statistics. Section 4 outlines research tasks that could benefit from the dataset and presents two preliminary experiments for two of these tasks. Finally, Section 5 concludes with a discussion of the potential impact of the presented dataset.

3 PATENTMATCH DATASET

The basis of our dataset is the *EP full-text data for text analytics* by the EPO.[1] It contains the XML-formatted full-texts and publication meta-data of all filed patent applications and published patent documents processed by the EPO since 1978. From 2012 onwards, the search reports for all patent applications are also included. In these reports, patent examiners cite paragraphs from prior art documents if these paragraphs are relevant for judging the novelty and inventive step of an application claim. Although there are no search reports available for applications filed before 2012, we do not discard these older applications because their corresponding published patent documents are frequently referenced as prior art. We use all available search reports to create a dataset of claims of

---

[1]https://www.epo.org/searching-for-patents/data/bulk-data-sets/text-analytics

Europäisches Patentamt European Patent Office Office européen des brevets	EUROPEAN SEARCH REPORT		Application Number EP 18 21 4053
DOCUMENTS CONSIDERED TO BE RELEVANT			
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
X Y	EP 1 351 172 A1 (SHIMA SEIKI MFG [JP]) 8 October 2003 (2003-10-08) * paragraph [0027] - paragraph [0028] * * paragraph [0060] * * paragraph [0070] - paragraph [0074] * -----	1,3-9 2	INV. G06K9/00 G06Q30/06 TECHNICAL FIELDS SEARCHED (IPC) G06K

Figure 1: In this excerpt from a search report, a patent examiner cites paragraph numbers of the published patent document *EP1351172A1* for assessing the novelty of claim 1 and 3-9 of application *EP18214053*.

patent applications matched with prior art, more precisely, paragraphs of cited *"X"* documents and *"A"* documents. Accordingly, *"X"* citations represent positive samples and *"A"* citations represent negative samples. These two categories *"X"* and *"A"* differ significantly regarding the level of semantic relevance of a given citation for a given claim. *"Y"* citations are not used in this work, as they seem too close to *"X"* citations with regard to their level of semantic relevance to generate a good training signal.

Our data processing pipeline uses Elasticsearch for storing and searching through this large corpus of about 210GB of text data. As a first data preparation step, an XML parser extracts the full text and meta-data from the raw, multi-nested XML files. Further, for each citation within a search report, it extracts claim number, patent application ID, date, paragraph number, and the type of the references, i.e., *"X"* document or *"A"* document.

Since the search reports were written in a rather systematic, but still unstructured and non-consistent way, a second parsing step standardizes the data format of paragraph references. References like "[paragraph 23]-[paragraph 28]" or "0023 - 28" are converted to complete enumerations of paragraph numbers "[23,24,25,26,27,28]". Furthermore, references by patent examiners comprise not only text paragraphs but also figures, figure captions, or the whole document. In our standardization process, all references that do not resolve to text paragraphs are discarded.

In the final step, we use the index of our Elasticsearch document database to resolve the referenced paragraph numbers (together with the corresponding document identifiers) to the paragraph texts. Similarly, we resolve the claim texts corresponding to the claim numbers. Thereby, we obtain a dataset that consists of a total of 6,259,703 samples, where each sample contains a claim text, a referenced paragraph text, and a label indicating one of the two types of reference: *"X"* document (positive sample) or *"A"* document (negative sample). Table 1 lists statistics of the full dataset and Figure 2 exemplifies a claim text and cited paragraph texts of positive and negative samples.

Table 1: Dataset statistics: Each sample is a pair of an application's claim and paragraph cited from either an *"X"* document (positive sample) or *"A"* document (negative sample).

Samples	6,259,703
"X" document citations	3,492,987
"A" document citations	2,766,716
Distinct patent applications	31,238
Distinct cited documents	33,195
Distinct claim texts	297,147
Distinct cited paragraphs	520,376
Median claim length (chars)	274
Median paragraph length (chars)	476

We also provide two variations of the data for simplified usage in machine learning scenarios. The first variation balances the label distributions by downsampling the majority class. For each sample with a claim text and a referenced paragraph labeled *"X"*, there is also a sample with the same claim text with a different referenced paragraph labeled *"A"* and vice versa. This balanced training set consists of 347,880 samples. In this version of the dataset, different claim texts can have different numbers of references. The number of *"X"* and *"A"* labels is only balanced for each claim text itself.

The second variation balances not only the label distribution but also the distribution of claim texts. Further downsampling ensures that there is exactly one sample with label *"X"* and one sample with label *"A"* for each claim text. As a result, every claim in the dataset occurs in exactly two samples. This restriction reduces the dataset to 25,340 samples.

The PATENTMATCH dataset is published online with example code that shows how to use it for supervised machine learning, and a description of the data collection and preparation process.[2] As the underlying raw data has been released by the EPO under Creative

---

Claim 1 of application EP17862550:
An engine for a ship, comprising: …an air supply apparatus supplying the air to the cylinder wherein the air supply apparatus includes an auxiliary air supply member …

Paragraphs 35-37 of "X" document US5271358A:
…the engine system 10 includes a second gaseous injector 57 in fluid communication with the cylinder bore 16 through fuel injection port 27 in addition to the gaseous fuel injector 56…

Paragraphs 31-32 of "A" document US2016298554A1:
…gaseous fuel may be injected from gaseous fuel injector 38 while the air intake ports 32 are open…

Figure 2: An excerpt from a search report showing a claim and cited paragraphs. The "X" document (positive sample) is novelty-destroying for the claim while the "A" document (negative sample) is not novelty-destroying and merely constitutes technical background.

Commons Attribution 4.0 International Public License, we also release our dataset under the same license.[3] To foster comparable evaluation settings in future work, we separated it into a training set (80%) and a test set (20%) with a time-wise split based on the application filing date: All applications contained in the training set have an earlier filing date than all applications contained in the test set (March 29th, 2017).

4 PRELIMINARY EXPERIMENTS

Modern information retrieval systems do not solely rely on matching keywords from queries with documents. Especially for complex information needs, semantic knowledge needs to be incorporated [5]. With the rise of deep learning models, as well as word and document embeddings, improvements in grasping the semantic meaning of queries and documents have been made [2]. A number of related tasks aim at finding semantically related information, making use of advanced semantic representations [6] and intelligent retrieval models [20]. Passage retrieval [13], document clustering [9], and question answering [28] all rely on identifying semantically related information.

Addressing a first exemplary task, we conducted preliminary experiments on text pair classification with Bidirectional Encoder Representations from Transformers (BERT) [4] as a baseline system. The text pair classification uses the same neural network architecture as the next sentence prediction task: Given a pair of sentences, the next sentence prediction task is to predict if the second sentence is a likely continuation of the first sentence. In our text pair classification scenario, given a claim text and a cited paragraph text, the task is to decide whether the paragraph corresponds to an "X" document (positive sample) or an "A" document (negative sample). To make this decision, the model needs to assess the novelty of the claim in comparison to the paragraph. To this end, it transforms the input text to sub-word tokens and transforms them to their embedding representations. These representation pass through 12

layers of bidirectional Transformers [26] and the final hidden state of the special token [CLS] encodes the output class label. Our implementation uses the FARM framework and the pre-trained bert-base-uncased model.[4]

The test set accuracy on the balanced variation of the data is 54%. On the second variation of the data, which contains exactly one "X" document citation and one "A" document citation per claim, the accuracy on the test set is 52%. For both variations, the accuracy improvements per training epoch are small and the validation loss stops to decrease after training for 6 epochs. It is not to our surprise that the task poses a difficult challenge and that a fine-tuned BERT model is only slightly better than random guessing. The complex linguistic patterns, the legal jargon, and the patent-domain-specific language make it sheer impossible for laymen to manually solve this task and therefore an interesting research challenge for future work.

A second exemplary task is dense passage retrieval (DPR). Inspired by the work by Karpukhin et al. [12], we transform the PATENTMATCH dataset into the DPR format used for open-domain question answering. Dense passage retrieval is the first step of open-domain question answering and the DPR format contains lists of questions, where each question is accompanied with the correct answer, a passage that contains the answer (positive context), and a passage that does not contain the answer but is still semantically similar to the question (hard negative context). We apply this format to our scenario of matching patent claims with passages from prior art, such that the claim represents the question and the paragraph text from the referenced "X" document is the positive context and the paragraph text from the referenced "A" document is the hard negative context. This version of the PATENTMATCH dataset contains exactly one sample with label "X" and one sample with label "A" for each claim text, which results in about 12500 triples (claim, positive, hard negative) in DPR format.

Using the dataset in DPR format, we train a DPR model, which comprises two BERT models (bert-base-uncased) [4]. One model encodes patent claims while the other encodes paragraph texts from "X" and "A" documents. As in the original DPR paper [12], we leverage in-batch negatives for training, which means that given a batch with claims and paragraph texts from corresponding "X" and "A" documents as positive and hard negative contexts, we use the positive context of each claim as an additional negative context for all other claims in the same batch. Using a batch size of 8, there are 8 claims in each batch, 8 positive contexts, 8 hard negative contexts, and implicitly also 7 in-batch (non-hard) negative contexts for each claim. The learning rate is set to 10^{-5} using Adam, linear scheduling with warm-up, and a dropout rate of 0.1. Due to memory constraints on the GPU, we limit the claim texts to 200 tokens and the paragraph texts to 256 tokens. In our preliminary experiment, the model achieves an average in-batch rank of 1.42 after training for 5 epochs, which means that the positive context is ranked between second and third position out of eight on average (rank 0 corresponds to first position). Although the method does not return perfect results, it is very useful as a tool for experts who now need to only look at a handful of candidates instead of thousands to find the right paragraph.

---

[3]https://creativecommons.org/licenses/by/4.0/

[4]https://github.com/deepset-ai/FARM, https://huggingface.co/bert-base-uncased

5 IMPACT & CONCLUSIONS

With this paper, we not only introduce an extensive dataset that can be used to train and test systems for the aforementioned tasks, but also provide training data for patent passage retrieval [21]: a very challenging search task mostly conducted by highly-trained patent-domain experts. The need to at least partially automate this task arises from the growing number of patent applications worldwide.

And with deep learning methods requiring large training sets, we hope to foster research in the patent analysis domain by providing such a dataset. We presented a novel dataset that comprises pairs of semantically similar texts in the patent domain. More precisely, the dataset contains claims from patent applications and paragraphs from prior art. It was created based on search reports by patent officers at the EPO. The simple structure of the dataset reduces the amount of patent-domain knowledge required for analyzing the data or using it for supervised machine learning. With the release of the dataset, we thus hope to foster research on the (semi-)automation of passage retrieval tasks and on user interfaces that support experts in searching through prior art and creating search reports.

Further, we hope to spark research in analysing how patent experts search for relevant patents and, maybe more interesting, which relevant patents they miss and for what reason. By providing the matched claims and paragraphs, the search process of patent officers can be analyzed and search results compared. For future work, our learned model could be used to adapt the experts' keyword queries for higher recall and to understand the relationship between results from manually curated queries and (relevant) results from deep learning models.

ACKNOWLEDGMENTS

REFERENCES

[1] L. Abdelgawad, P. Kluegl, E. Genc, S. Falkner, and F. Hutter. Optimizing neural networks for patent classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 688–703, 2019.

[2] D. Cohen and W. B. Croft. A hybrid embedding approach to noisy answer passage retrieval. In *Advances in Information Retrieval*, pages 127–140, 2018.

[3] D. Cohen, L. Yang, and W. B. Croft. WikiPassageQA: A benchmark collection for research on non-factoid answer passage retrieval. In *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*, page 1165–1168, 2018.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, pages 1–16, 2018.

[5] M. Fernández, I. Cantador, V. López, D. Vallet, P. Castells, and E. Motta. Semantically enhanced information retrieval: An ontology-based approach. *Journal of Web Semantics*, 9(4):434–452, 2011.

[6] D. Ganguly, D. Roy, M. Mitra, and G. J. Jones. Word embedding based generalized language model for information retrieval. In *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*, pages

[7] J. Gobeill and P. Ruch. Bitem site report for the claims to passage task in CLEF-IP 2012. In *Proceedings of the CLEF-IP Workshop*, 2012.

[8] M. Golestan Far, S. Sanner, M. R. Bouadjenek, G. Ferraro, and D. Hawking. On term selection techniques for patent prior art search. In *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 803–806, 2015.

[9] A. Huang. Similarity measures for text document clustering. In *Proceedings of the New Zealand Computer Science Research Student Conference (NZCSRSC)*, volume 4, pages 9–56, 2008.

[10] J. A. Jeffery. Preserving the presumption of patent validity: An alternative to outsourcing the us patent examiner's prior art search. *Cath. UL Rev.*, 52:761, 2002.

[11] A. Judea, H. Schütze, and S. Brügmann. Unsupervised training set generation for automatic acquisition of technical terminology in patents. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 290–300, 2014.

[12] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020.

[13] M. Kaszkiel and J. Zobel. Passage retrieval revisited. In *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 178–185, 1997.

[14] T. Kulahcioglu, D. Fradkin, and S. Palanivelu. Incorporating task analysis in the design of a tool for a complex and exploratory search task. In *Proceedings of the Conference on Conference Human Information Interaction and Retrieval (CHIIR)*, page 373–376, 2017.

[15] Z. Lei and B. D. Wright. Why weak patents? testing the examiner ignorance hypothesis. *Journal of Public Economics*, 148:43 – 56, 2017. ISSN 0047-2727. doi: https://doi.org/10.1016/j.jpubeco.2017.02.004. URL http://www.sciencedirect.com/science/article/pii/S0047272717300178.

[16] S. Li, J. Hu, Y. Cui, and J. Hu. Deeppatent: patent classification with convolutional neural networks and word embedding. *Scientometrics*, 117(2):721–744, 2018.

[17] K. Loveniers. How to interpret epo search reports. *World Patent Information*, 54: 23–28, 2018.

[18] E. Marttin and A.-C. Derrien. How to apply examiner search strategies in espacenet. a case study. *World Patent Information*, 54:S33 – S43, 2018. ISSN 0172-2190. doi: https://doi.org/10.1016/j.wpi.2017.06.001. URL http://www.sciencedirect.com/science/article/pii/S0172219016301089. Best of Search Matters.

[19] J. Michel and B. Bettels. Patent citation analysis. a closer look at the basic input data from patent search reports. *Scientometrics*, 51(1):185–201, 2001.

[20] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, and R. Ward. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):694–707, 2016.

[21] F. Piroi, M. Lupu, A. Hanbury, A. P. Sexton, W. Magdy, and I. V. Filippov. CLEF-IP 2012: Retrieval experiments in the intellectual property domain. In *Proceedings of the CLEF-IP Workshop*, pages 1–16, 2012.

[22] J. Risch and R. Krestel. Domain-specific word embeddings for patent classification. *Data Technologies and Applications*, 53(1):108–122, 2019.

[23] T. Russell-Rose, J. Chamberlain, and F. Shokraneh. A visual approach to query formulation for systematic search. In *Proceedings of the Conference on Human Information Interaction and Retrieval (CHIIR)*, page 379–383, 2019.

[24] S. Suzuki and H. Takatsuka. Extraction of keywords of novelties from patent claims. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 1192–1200, 2016.

[25] Y.-H. Tseng and Y.-J. Wu. A study of search tactics for patentability search: A case study on patent engineers. In *Proceedings of the Workshop on Patent Information Retrieval (PaIR@CIKM)*, pages 33–36, 2008.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.

[27] X. Xue and W. B. Croft. Transforming patents into prior-art queries. In *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 808–809, 2009.

[28] S. Yang, L. Zou, Z. Wang, J. Yan, and J.-R. Wen. Efficiently answering technical questions—a knowledge graph approach. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, pages 3111–3118, 2017.

795–798, 2015.