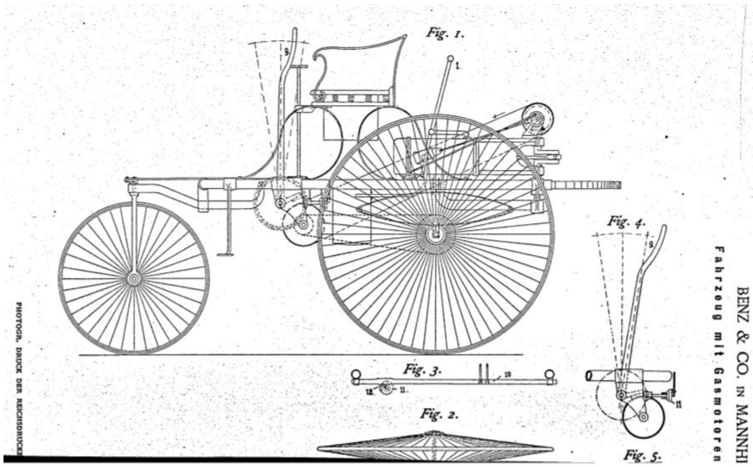
Proceedings of the

3rd Workshop on Patent Text Mining and Semantic Technologies

Editors: Ralf Krestel
Hidir Aras
Linda Andersson
Florina Piroi
Allan Hanbury
Dean Alderucci



PatentSemTech 2022

Madrid, Spain
July 2022

Editors:

Ralf Krestel, r.krestel@zbw.eu
ZBW - Leibniz Information Centre for Economics & Kiel University,
Kiel, Germany

Hidir Aras, Hidir.Aras@fiz-Karlsruhe.de
Text and Data Mining (TDM), Leibniz Institute for Information Infrastructure,
FIZ Karlsruhe GmbH, Karlsruhe, Germany

Linda Andersson, Linda.Andersson@artificialresearcher.com
Artificial Researcher IT, GmbH, Vienna, Austria

Florina Piroi, Florina.Piroi@tuwien.ac.at
Institute of Information Systems Engineering, Technische Universität Wien,
Vienna, Austria
Research Studio Austria, Data Science Studio, Vienna, Austria

Allan Hanbury, Allan.Hanbury@tuwien.ac.at
Institute of Information Systems Engineering, Technische Universität Wien,
Vienna, Austria

Dean Alderucci, dalderuc@cs.cmu.edu
Center for AI and Patent Analysis, Carnegie Mellon University, Pittsburgh, PA, USA

This document contains the informal proceedings of the 3rd Workshop on on Patent Text Mining and Semantic Technologies (PatentSemTech 2022) held on July 15, 2019 in Madrid, Spain. All submissions to this workshop have gone through single-blind reviewing process, to assess their relevance to the workshop topics.

The workshop was organized as a one day event co-located with the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, that took place in July 11-15, 2022, in Madrid, Spain.

The workshop has been partially supported by the DoSSIER Project with funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860721

DOI: 10.34726/3550

Contents

Foreword

The third edition of the *Patent Text Mining and Semantic Technologies* (PatentSemTech'22) workshop series was held as a full-day hybrid event in conjunction with the SIGIR 2022 conference. The workshop focused on research and new developments from relevant fields such as Natural Language Processing, Text and Data Mining and Semantic Technologies applied to Patent Retrieval and Patent Analytics.

The workshop's main aim was to address the adaptation of existing Natural Language Processing (NLP) methods, Machine Learning and Deep Learning-based tools (ML/DL) to the search and analytics of patent data. The challenge in this domain is that patent data is complex in content, contains lengthy documents, and provides a heterogeneous type of scientific text that covers diverse scientific subject areas, like chemistry, pharmacology, engineering, communication technologies, etc. Therefore, patent data is, compared to general language text corpora, more difficult to analyse. Processing patent data has multiple facets that should be exploited to obtain good results:

- It constitutes a huge corpus of scientific-technical documents for a variety of technological domains,
- They are rich in available meta-data such as spatial data, bibliographic data, classifications, temporal data, etc.
- Patents describe essential scientific-technical knowledge enclosing solutions for real-world applications,
- They are complementary knowledge to scientific literature, e.g. chemical and physical properties, bio-science knowledge for drug-target-interaction, which appears first in patents, mostly not published elsewhere

During the full-day event at SIGIR 2023, a number of 2-4 page contributions were presented and discussed. The contributions were screened for relevancy and quality by the workshop organizers. The workshop included a keynote contribution given by Henry (Jamie) Holcombe, Chief Information Officer (CIO) at the United States Patent and Trademark Office (USPTO), who presented the Office's long term strategies for managing the large number of patent documents in an efficient way. The workshop concluded with a lively discussion round where all the participants expressed their ideas and possible solutions to the challenges of working with patent data.

Germany, Austria, USA, July 2022

Ralf Krestel,
Hidir Aras,
Linda Andersson,
Florina Piroi,
Allan Hanbury,
Dean Alderucci

End-to-End Chemical Reaction Extraction from Patents

Yuan Li
yuan.li1@unimelb.edu.au
The University of Melbourne
Melbourne, Australia

Jiayuan He*
jiayuan.he@rmit.edu.au
RMIT University
Melbourne, Australia

Hiyori Yoshikawa*
hiyori.yoshikawa@unimelb.edu.au
Fujitsu Limited
Minato Ward, Japan

Biaoyan Fang
biaoyanf@student.unimelb.edu.au
The University of Melbourne
Melbourne, Australia

Zenan Zhai
zenan.zhai@student.unimelb.edu.au
The University of Melbourne
Melbourne, Australia

Christian Druckenbrodt
c.druckenbrodt@elsevier.com
Elsevier Information Systems GmbH
Frankfurt, Germany

Camilo Thorne
c.thorne.1@elsevier.com
Elsevier Information Systems GmbH
Frankfurt, Germany

Saber A. Akhondi
s.akhondi@elsevier.com
Elsevier BV
Amsterdam, Netherlands

Karin Verspoor*†
karin.verspoor@rmit.edu.au
RMIT University
Melbourne, Australia

ABSTRACT

With the rapid growth of chemical patents, there is increasing demand for automated extraction of information relating to chemical compounds and their synthesis from patents. Although there are existing models that can extract chemical entities and reaction events, these have significant practical limitations. First, they typically cannot process a full patent document, targeting short texts containing only reaction descriptions. Second, they neglect reaction texts where steps in the reaction are elided through reference to other reactions. To address these issues, we propose an integrated and comprehensive chemical reaction extraction system consisting of a pipeline of components for reaction detection, chemical named entity recognition, event extraction, anaphora resolution, reaction reference resolution, and table classification.

CCS CONCEPTS

• **Computing methodologies → Information extraction**.

KEYWORDS

information extraction, named entity recognition, event extraction, anaphora resolution, chemical reactions, patent text mining

1 INTRODUCTION

The discovery of new chemical compounds is a key driver of the chemistry and pharmaceutical industries, *inter alia*. Patents serve as a critical source of information about new chemical compounds, providing timely and comprehensive information about new chemical compounds [1, 2]. Despite the significant commercial and research value of the information in patents, manual effort is still the primary mechanism for extracting and organizing this information. This is costly, considering the large volume of patents available [11]. Development of automatic natural language processing (NLP) systems for chemical patents, which aim to convert text corpora into structured knowledge about chemical compounds, has become a focus of recent research [9, 10].

In this study we consider a system that focuses on chemical reaction processes described in chemical patents. A chemical reaction is a process leading to the transformation of one set of chemical substances to another. A full reaction requires at least the starting materials and the final product to be defined, and usually includes information such as reagents, catalysts, and experiment conditions to further describe the reaction. Our overarching objective is to enable the automatic identification of each reaction described in a complete patent document, and to fully characterize each reaction by extracting each relevant component.

2 SYSTEM OVERVIEW

To perform end-to-end extraction of chemical reactions from full patents, we define a pipeline of interconnected NLP tasks.

Reaction snippet detection: We first need to locate reaction descriptions in a patent, for processing in downstream tasks. We formulate this task as a paragraph-level sequence tagging problem, where a patent is given as a sequence of paragraphs and the task is to detect a span of contiguous paragraphs describing a single chemical reaction. We train a BiLSTM-CRF model for this task on the dataset described in [13] using the same experimental settings.

Chemical NER:. Using the reaction snippets extracted from full patents, the task to identify chemical entities and their roles in a chemical reaction can be formulated as named entity recognition

*Also with The University of Melbourne.
†Corresponding author.

Published in "Proceedings of The 3rd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2022)", Eds. H. Aras, R. Krestel, F. Piroi, L. Andersson, A. Hanbury, D. Alderucci, PatentSemTech 2022, 15 July 2022, Madrid, Spain. https://doi.org/10.34726/3550 | © Authors 2022. CC BY-NC-ND 4.0 License..

(NER). We train a BERT-CRF model for this task using the annotation schema and data for chemical NER task detailed in [7, 8].

Event extraction: A chemical reaction usually consists of an ordered sequence of *event steps* that either transforms a starting material into a product or just purifies or isolates a chemical substance. An event is characterised by (a) a trigger word that flags its occurrence, and (b) a relation connecting the trigger word and chemical entities involved in the event. For this task, we use a BERT-CRF model to extract trigger words and chemical entities from snippets and borrow ideas from the span-based BERT model in [5]. In this approach, all pairs of trigger words and entities are enumerated, BERT is applied to obtain the contextualized representation of each relevant token, and a classifier decides the nature of the relation between them using pooling of token representations.

Anaphora resolution: There are rich anaphoric relations *between* and *within* event steps. We consider two main types of anaphoric relations defined in [6]: coreference, where two mentions refer to the same entity, and bridging, linking a chemical compound and its source. We decompose this task into (a) anaphor mention detection and (b) relation classification. We use a BERT-CRF model for mention detection. For relation classification, we adopt the span-based BERT model proposed in [4].

Reaction reference resolution: So far, we have assumed that a reaction snippet contains the complete information of a chemical reaction. However, chemical patents often detail several similar compounds that have a common substructure and can be synthesized in analogous ways. They contain many references connecting descriptions of similar chemical reactions, to avoid redundancy in describing common reaction conditions. This leads to the problem of identifying references from an incomplete snippet to others. Here, we use the model proposed in [12], first determining if a snippet has others that refer to it, and then enumerating possible reference pairs of snippets and classifying them.

Table classification: Apart from text paragraphs, a large amount of information in patents is represented in tables and images. Here, we focus on identifying tables containing chemical reaction properties such as starting materials, products, yields, etc. To differentiate tables of interest from others, we train a Table-BERT classifier [3] on the ChemTables data [14]. The model first concatenates all tokens within all cells from the table and then takes the flattened table as input. For tables classified into reaction properties category, we further extract reactions based on the table header if there are sufficient information describing reactions.

3 DISCUSSION

We have introduced the essential requirements for building a comprehensive chemical reaction extraction system covering a wide range of tasks. We have proposed an initial approach for each step leveraging existing data resources from the ChEMU shared tasks, illustrating how the individual tasks can be brought together into a coherent whole. This integration addresses two key limitations of previous studies: our system can process full patent documents directly, and we can find the snippets an incomplete reaction snippet refers to. We leave performance evaluation of individual steps, as well as the complete system, to a more in-depth presentation. In the future, we plan to further develop this framework to extract complete reaction information by incorporating inference over reaction references, and to extend the scope of our system to handle images and chemical structures. Opportunities also exist to explore joint modelling or multi-task learning across the constituent tasks in this pipeline, for instance coupling NER and anaphora resolution.

REFERENCES

[1] Saber A Akhondi, Hinnerk Rey, Markus Schwörer, Michael Maier, John Toomey, Heike Nau, Gabriele Ilchmann, Mark Sheehan, Matthias Irmer, Claudia Bobach, et al. 2019. Automatic identification of relevant chemical compounds from patents. *Database* 2019 (2019).

1

[2] Mervyn Bregonje. 2005. Patents: A unique source for scientific technical information in chemistry related industry? *World Patent Information* 27, 4 (2005), 309–315.

[3] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. TabFact: A Large-scale Dataset for Table-based Fact Verification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.* OpenReview.net.

[4] Ritam Dutt, Sopan Khosla, and Carolyn P. Rosé. 2021. A pipelined approach to Anaphora Resolution in Chemical Patents. In *Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum (CEUR Workshop Proceedings, Vol. 2936).* CEUR-WS.org, 710–719.

[5] Markus Eberts and Adrian Ulges. 2020. Span-Based Joint Entity and Relation Extraction with Transformer Pre-Training. In *ECAI 2020 - 24th European Conference on Artificial Intelligence (Frontiers in Artificial Intelligence and Applications, Vol. 325).* IOS Press, 2006–2013.

[6] Biaoyan Fang, Christian Druckenbrodt, Saber A. Akhondi, Jiayuan He, Timothy Baldwin, and Karin M. Verspoor. 2021. ChEMU-Ref: A Corpus for Modeling Anaphora Resolution in the Chemical Domain. In *Proc. 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021.* Association for Computational Linguistics, 1362–1375.

[7] Jiayuan He, Biaoyan Fang, Hiyori Yoshikawa, Yuan Li, Saber A. Akhondi, Christian Druckenbrodt, Camilo Thorne, Zubair Afzal, Zenan Zhai, Lawrence Cavedon, Trevor Cohn, Timothy Baldwin, and Karin Verspoor. 2021. ChEMU 2021: Reaction Reference Resolution and Anaphora Resolution in Chemical Patents. In *Advances in Information Retrieval - 43rd European Conf. on IR Research, ECIR 2021, Part II (Lecture Notes in Computer Science, Vol. 12657).* Springer, 608–615.

[8] Jiayuan He, Dat Quoc Nguyen, Saber A. Akhondi, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Zubair Afzal, Zenan Zhai, Biaoyan Fang, Hiyori Yoshikawa, Ameer Albahem, Lawrence Cavedon, Trevor Cohn, Timothy Baldwin, and Karin Verspoor. 2020. Overview of ChEMU 2020: Named Entity Recognition and Event Extraction of Chemical Reactions from Patents. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th Intl Conf of the CLEF Association, CLEF 2020 (Lecture Notes in Computer Science, Vol. 12260).* Springer, 237–254.

[9] Jiayuan He, Dat Quoc Nguyen, Saber A. Akhondi, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Zubair Afzal, Zenan Zhai, Biaoyan Fang, Hiyori Yoshikawa, Ameer Albahem, Lawrence Cavedon, Trevor Cohn, Timothy Baldwin, and Karin Verspoor. 2021. ChEMU 2020: Natural Language Processing Methods Are Effective for Information Extraction From Chemical Patents. *Frontiers Res. Metrics Anal.* 6 (2021), 654438.

[10] Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. 2015. CHEMDNER: The drugs and chemical names extraction challenge. *Journal of Cheminformatics* 7, 1 (2015), 1–11.

[11] Sorel Muresan, Plamen Petrov, Christopher Southan, Magnus J Kjellberg, Thierry Kogej, Christian Tyrchan, Peter Varkonyi, and Paul Hongxing Xie. 2011. Making every SAR point count: the development of Chemistry Connect for the large-scale integration of structure and bioactivity data. *Drug Discovery Today* 16, 23-24 (2011), 1019–1030.

[12] Hiyori Yoshikawa, Saber Akhondi, Camilo Thorne, Christian Druckenbrodt, Ralph Hoessel, Zenan Zhai, Jiayuan He, Timothy Baldwin, and Karin Verspoor. 2021. Chemical Reaction Reference Resolution in Patents. In *Proc. 2nd Workshop on Patent Text Mining and Semantic Technologies.*

[13] Hiyori Yoshikawa, Dat Quoc Nguyen, Zenan Zhai, Christian Druckenbrodt, Camilo Thorne, Saber A. Akhondi, Timothy Baldwin, and Karin Verspoor. 2019. Detecting Chemical Reactions in Patents. In *Proc. 17th Annual Workshop of the Australasian Language Technology Association, ALTA 2019, Sydney, Australia, December 4-6, 2019.* 100–110.

[14] Zenan Zhai, Christian Druckenbrodt, Camilo Thorne, Saber A. Akhondi, Dat Quoc Nguyen, Trevor Cohn, and Karin M. Verspoor. 2021. ChemTables: A dataset for semantic classification on tables in chemical patents. *J. Cheminformatics* 13, 1 (2021), 97.

Optimizing BERT-based reference mining from patents

Zahra Abbasiantaeb
z.abbasiantaeb@sbb.leidenuniv.nl
Leiden Institute of Advanced Computer Science
Leiden, the Netherlands

Suzan Verberne
s.verberne@liacs.leidenuniv.nl
Leiden Institute of Advanced Computer Science
Leiden, the Netherlands

Jian Wang
j.wang@sbb.leidenuniv.nl
Leiden Institute of Advanced Computer Science
Leiden, the Netherlands

CCS CONCEPTS

• **Information systems → Information extraction**.

KEYWORDS

reference mining, citations, patents, sequence labelling, BERT

ACM Reference Format:
Zahra Abbasiantaeb, Suzan Verberne, and Jian Wang. 2022. Optimizing BERT-based reference mining from patents. In *Proceedings of the 3rd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech) at SIGIR 2022.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Extracting references from the unstructured text of patents is of great importance for those who study the link between science and technology. The scientific references (citations) in patents provide valuable information for investigating the impact of science on technological advances. For example, the features of a scientific paper that would lead to technological advances can be extracted.

We divided the task of identifying patent–science references into two main steps namely (1) reference extraction and (2) reference matching. In the first step, we need to extract the string of the references from the unstructured text of the patents and in the second step we need to match the extracted reference string to a database of scientific papers. In the first step, the key point is extracting all of the in-text references from the patent's text. Consequently, improving the recall is of a higher value than precision because the wrong extracted references would be omitted in the second step when they do not match with any publication.

In this work, we are going to improve the BERT-based reference extraction model from prior work [7] by (1) using multiple pre-trained BERT models, including patent-specific models, (2) using a more effective method for sequence splitting, (3) investigating the impact of down sampling on our model to cope with the class imbalance, (4) and annotating a larger dataset for training the model on the reference extraction task. We will collect a large size dataset that links the in-text references of the USPTO and EPO patents dataset to scientific publications. In this extended abstract we present our on-going work, including the results for step (1)-(3) and our plans for step (4).

2 OUR REFERENCE EXTRACTION METHOD

We consider the reference extraction problem as a sequence labeling task, following prior work [6, 7]. The goal of reference extraction is to predict the beginning and inside tokens of a reference in the input patent text. To this aim, we utilize pre-trained language models like BERT-base [2], SciBERT [1], PatentBERT [3], BERT for patents [5], and BioBERT [4]. We added an additional classification layer on top of these models for classifying each token as 'I' (inside of a reference), 'B' (beginning of a reference), and 'O' (outside of reference). The PatentBERT model is generated by fine-tuning the BERT-base model on the claims of the USPTO patent dataset for the patent classification task. While the *BERT for patents* model is generated by further pre-training the BERT-large model on the whole text of patents including description, abstract, and claims.

Sequence splitting. Having the texts of patents, we generate the samples by segmenting the text into a sequence of n tokens. The number n is calculated in a way that ensures the length of sub-tokens will not exceed the maximum length for the model's input. We selected the maximum length of 512 for the model's input in our experiments because this is the maximum sequence length in BERT. Compared to prior work, which used sentence splitting [7], our longer sequences provide more context information for the model. First, we tokenize the whole patent text, using the tokenizer of the corresponding pre-trained model to obtain the number of sub-tokens ($|t|$) for each token t. Then, we select each sequence as follows:

$$\{t_1, t_2, t_3, ..., t_n\} : \max_n |t_1| + |t_2| + |t_3| + ... + |t_n| <= 512 \quad (1)$$

Downsampling. As the occurrence of references in patent texts is relatively sparse, we have many sequences with no references. We used down sampling to remove the sequences with no reference in our train set. With down sampling, the models train faster and more robustly. Our experiments reveal a boost in the performance of the model with down sampling.

Data. We use the annotated 22 patents dataset collected by [6] from Google Patents. All patents in this sample have IPC class C12N and were published in 2010. Each token in this dataset was hand-labeled with IOB labels. The whole dataset includes 2,318 'B' tokens (thus 2,318 references) and 32,359 'I' tokens. After segmentation in sequences, we converted the dataset of 22 patents into 14,270 sequences where 8,530 of them had no 'B' or 'I' labels.

Published in "Proceedings of The 3rd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2022)", Eds. H. Aras, R. Krestel, F. Piroi, L. Andersson, A. Hanbury, D. Alderucci, PatentSemTech 2022, 15 July 2022, Madrid, Spain. https://doi.org/10.34726/3550 |

Table 1: Results of experiments on effect of down sampling using BERT-base on 22 patents dataset.

Down Sampling	Label	Precision	Recall
True	B	0.884	0.922
	I	0.967	0.966
False	B	0.880	0.919
	I	0.959	0.964

3 EXPERIMENTS AND RESULTS

We evaluated our models using Leave-One-Out Cross-Validation (LOOCV). We implemented our models using the HuggingFace framework (here is our source code[1]). We train our models for six epochs with the learning rate of 10^{-4}.

The effect of downsampling. The results of this experiment are shown in Table 1. The table shows that down sampling results in higher recall and precision. By applying down sampling, the model trained more smoothly. The sequences with no reference are less informative for the model and the long sequences with 512 tokens, provide enough context for the model to recognize the references.

Comparing BERT models. The result of these experiments is shown in Table 2. We used SciBERT, BioBERT, and PatentBERT models which are based on BERT-base while the BERT for patents model is based on the BERT-Large model and it was further pre-trained on patent data. All of the models except the PatentBERT and the BERT for patents model were 'cased', which means that upper- and lowercase has been retained on those models. As shown in the tabel, SciBERT outperforms the other models in the recall and *BERT for patents* outperforms the rest of the models in precision. The superb performance of the SciBERT model can be due to the fact that SciBERT is also fine-tuned with sequence tagging tasks on scientific texts. The Patent-BERT model does not outperform the SciBERT and the BERT for patents models. This can be because it is only fine-tuned on the claims of the patents and the claims do not include any references. In addition, for the in-text reference extraction task the case of the tokens is a very informative. This fact can explain the lower performance of Patent-BERT.

Finally, we observe that our BERT for patents model outperforms the previous work on the 22 patents dataset with 1.4% point on 'B' precision and our SciBERT model outperforms the baseline with 1.4% and 1.0% on 'B' recall and 'I' recall, respectively.

Table 2: Results of evaluation on 22 patents dataset. The baseline SciBERT model was trained without down sampling and on shorter sequence lengths (sentences).

Model	Label	Precision	Recall
BERT-base	B	0.884	0.922
	I	0.967	0.966
SciBERT (base)	B	0.954	**0.968**
	I	0.980	**0.986**
BERT for patents (large)	B	**0.961**	0.965
	I	0.983	0.978
Patent-BERT (base)	B	0.945	0.963
	I	0.979	0.970
BioBERT (base)	B	0.952	0.962
	I	0.984	0.980
Our Baseline (SciBERT) [7]	B	0.947	0.954
	I	0.986	0.976

4 DISCUSSION AND FUTURE WORK

The results of our model, as shown in Table 2, depicts almost perfect scores (Recall is 96.8% for B-labels and 98.6% for I-labels). However, this great improvement and results are based on small scale evaluation on a single-domain sample. In order to have a large scale and complete evaluation, we plan to further improve our reference extraction model using a larger and more diverse set of training dataset. We draw a random sample of (between 500 and 1000 patents) EPO and USPTO patents from all technological fields. We have hired student assistants to read through the patent full texts and manually label in-text reference strings. We have recruited 8 students for 4 weeks in June and 10 hours per week. This dataset may help us to further improve and evaluate our reference extraction model. In addition, we will investigate the impact of considering a special token for the end token of references rather than labeling them as 'I'.

With a model trained on the larger data set we will extract references from the complete EPO and USPTO collections. Next, we will focus on matching the extracted references to a publication database (the Web of Science). For the reference matching task, we will improve the available model [6] by mitigating its weaknesses, in particular the problem of ambiguous matching. For example, in the case that we only have the name of the author and year of publication, we can use text matching models to match the abstract of the patent with the title of all of the author's publications in that year, to find the most relevant publication. Finally, we intend to further pre-train a BERT cased model on a large patent collection, to be used for sequence labelling tasks in the patent domain.

REFERENCES

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: a pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, (Nov. 2019), 3615–3620. doi: 10.18653/v1/D19-1371.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, (June 2019), 4171–4186. doi: 10.18653/v1/N19-1423.

[3] Jieh-Sheng Lee and Jieh Hsiang. 2019. Patentbert: patent classification with fine-tuning a pre-trained bert model. *arXiv preprint arXiv:1906.02124*.

[4] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36, 4, 1234–1240.

[5] J. Yonamine R. Srebrovic. 2020. Leveraging the bert algorithm for patents with tensorflow and bigquery. (2020). https://services.google.com/fh/files/blogs/bert_for_patents_white_paper.pdf.

[6] Suzan Verberne, Ioannis Chios, and Jian Wang. 2019. Extracting and matching patent in-text references to scientific publications. In *BIRNDL@ SIGIR*, 56–69.

[7] K Voskuil and S Verberne. 2021. Improving reference mining in patents with bert. In *Proceedings of the 11th International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2021)*, 78–88.

[1] https://github.com/ZahraAbbasiantaeb/Patent-in-text-reference-extraction

An Ensemble Architecture of Classifiers for Patent Classification

Eleni Kamateri
International Hellenic University
Thessaloniki 57400, Greece
ekamater@hotmail.com

Michail Salampasis
International Hellenic University
Thessaloniki 57400, Greece
msa@ihu.gr

CCS CONCEPTS

• Social and professional topics -> Computing / technology policy -> Intellectual property -> Patents • Computing methodologies -> Machine learning -> Learning paradigms -> Supervised learning -> Supervised learning by classification

KEYWORDS

Patent, Classification, Single-label, Sub-classes, Groups, Bi-LSTM

INTRODUCTION

One important task when a patent application is submitted is to assign one or more classification codes. Correct pre-classification will enable routing of the patent application to an appropriate patent examiner who is knowledgeable of the specific technical field. This task is undertaken by patent office desks, however due to the large number of applications and the potential complexity of an invention, they are usually overwhelmed. Additionally, in several innovation related tasks it is important to identify the technical area of an idea or potential invention as it is represented in the international patent classification taxonomy. Therefore, there is a need to support this manual task or even to fully automate it, hopefully with an accuracy close to patent professionals.

Patent classification can be a single- or multi-label text classification problem dealing with patent documents that are long technical documents having a quite distinctive language and structure. Moreover, classification schemes used to classify patent documents follow a hierarchical structure containing thousands of codes each representing a more general (at higher levels) or a very specific (at lower levels) technological concept.

Research efforts in this field have tried to automate the patent classification task [1-4], bringing together NLP and ML/DL techniques for efficient patent modelling and representation and automatic classification. However, they failed to achieve high performance although they applied various simplifications, e.g., working with well-represented codes having many training samples or working at higher level of the classification hierarchy.

A promising ML method that can improve the performance of learning models are ensemble techniques, which combine the results from multiple models. An ensemble technique receives evidence from multiple learning models, working either at the same or different sources of information, combines these evidences to produce improved results, i.e. more accurate predictions than a single model would [5]. The ensemble techniques exploit potentially not related information coming from all single models involved and this is the reason they attain better performance.

Although ensemble techniques produced good results in many applications, they have been less explored for automated patent classification. One of these examples is presented in [6] where a variety of combination techniques of different ML methods (kNN, LLSF, NN, Winnow) was explored to improve the overall performance. Moreover, in [7], the proposed ensemble technique is performed only at the upper levels of the IPC hierarchy. Last, Kamateri et al. [8] presented an ensemble technique, which consists of three identical individual classifiers (CNN, bi/GRU, bi/LSTM) with each of them trained on a different part of the patent text, i.e., the title-abstract, the description and the claims section, respectively, obtaining better results than each of those classifiers acting on its own.

In this study, we extend the work already presented in [8] and introduce a new ensemble architecture for automated patent classification at multiple levels. The ensemble architecture is instantiated in the single-label pre-classification task at the subclass (3rd) and group (4th) level category of the IPC 5+ level hierarchy. Our first results are quite promising showing that the combination of classifiers significantly outperforms the same classifiers when used as standalone solutions as well as the current state-of-the-art techniques.

ENSEMBLE ARCHITECTURE

An ensemble architecture (Fig. 1) can consist of individual classifiers that can be of any number and type, while they can be trained with the same or different parts of the patent document. Each classifier produces a list of probabilities for all labels based on its whole or partial knowledge about the patent. Then, the probabilities for a specific label derived from all individual classifiers are combined and the final probability is calculated for this label. The label with the maximum probability consists the predicted label for the patent. The combination of probabilities of the individual classifiers can be aggregated using simple/weighted averaging, voting, stacking or other combination techniques.

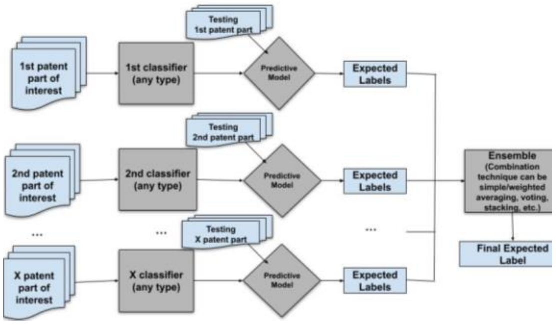

Testing
1st patent part
1st patent part of interest
1st classifier (any type)
Predictive Model
Expected Labels
Testing
2nd patent part
2nd patent part of interest
2nd classifier (any type)
Predictive Model
Expected Labels
Ensemble (Combination technique can be simple/weighted averaging, voting, stacking, etc.)
...
...
...
Testing
X patent part
X patent part of interest
X classifier (any type)
Predictive Model
Expected Labels
Final Expected Label

Figure 1: Ensemble architecture of classifiers

Published in "Proceedings of The 3rd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2022)", Eds. H. Aras, R. Krestel, F. Piroi, L. Andersson, A. Hanbury, D. Alderucci, PatentSemTech 2022, 15 July 2022, Madrid, Spain. https://doi.org/10.34726/3550 | © Authors 2022. CC BY-NC-ND 4.0 License.

		1. Abstract	2. Description	3. Claims	4. Title	5. Applicants	6. Inventors
Subclass /Group	Individual Bi-LSTM classifier	63.76/44.68	66.46/47.23	64.56/45.10	59.58/40.74	24.32/12.93	11.52/6.01
	Ensemble (simple averaging of classifiers 1-6)	70.67/53.06					
	Ensemble (weighted averaging of classifiers 1-6)	70.70/53.11					

Table 1: Accuracy at subclass/group level

DATA COLLECTION

The CLEF-IP 2011 test collection was imported in a MySQL database and the latest version of English patent documents which contain the required information were used for evaluating the proposed technique. The code for re-creating the MySQL database and the specific dataset used in this study are available online[1].

EXPERIMENTS

An ensemble of bidirectional LSTM classifiers was employed, since this ML method has been proved in [8] to attain better results than other DL methods. Each classifier was trained on a particular patent part, i.e., title, abstract, description, claims, inventors and applicants, respectively. The outcome probabilities of individual classifiers were aggregated using simple and weighted averaging.

With respect to the patent representation, the first 60 words from the patent part of interest (e.g., title, abstract, etc.) were used after undertaking a sequence of preprocessing steps (cleaning punctuation, symbols and numbers, and stop word removal). The feature words were then mapped to embeddings using a domain-specific pre-trained language model which has been created on a patent dataset, proposed by Risch and Krestel [4].

The dataset of 541,131 patents was split into training, validation and testing sets (80:10:10). Batch size was set to 128 and epochs to 15.

RESULTS

Table 1 presents the *Accuracy* of six individual classifiers trained on a specific patent part and the *Accuracy* of the ensemble technique. The accuracy is much improved when an ensemble technique is applied combining (with weighted averaging) the predicted probabilities acquired by individual classifiers working on a specific patent part. Moreover, the weighted averaging seems to slightly outperformed simple averaging. It is also clear that the proposed method provides better results than those obtained from current state-of-the-art techniques (e.g., an accuracy of 67% for subclass level using LSTM and Word2Vec trained on patent data [9] and 36.89% for group level [10]).

ANALYSIS OF RESULTS

Looking carefully at the experimental results, we observe that under-represented codes are those that mainly affect the accuracy score (Fig. 2). The usage of a more balanced dataset excluding such codes further enhances the accuracy scores indicating the general expectation of any ML method that larger quantities of training data will produce better results. To what extent and up to which point, increasing training data will enhance the performance it remains an open research question.

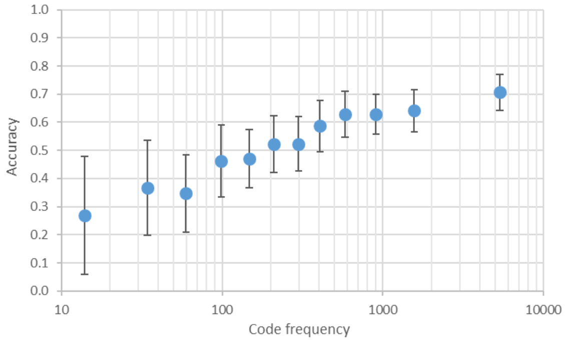


Figure 2: Accuracy as a function of average code frequency for groups of 50 subsequent codes; error bars denote σ.

Another interesting note is that this ensemble of classifiers seems to achieve better performance when it exploits not related information. For example, when we tried to combine the predicted probabilities from two classifiers, the first trained with first 60 words of a patent description (achieving an accuracy of 66.46%) and the second trained with the subsequent 60 words (achieving an accuracy of 56.15%) then the outcome performance was slightly improved (67.23%/67.66% for simple/weighted averaging).

CONCLUSIONS

In this study, an ensemble architecture is presented to address the automated patent classification problem at multiple levels. An ensemble architecture of bidirectional LSTM classifiers was employed in the single-label pre-classification task getting an accuracy of 70.70% at the subclass and 53.11% at the group level.

REFERENCES

[1] Grawe, M. F., Martins, C. A., & Bonfante, A. G. (2017). Automated patent classification using word embedding. *In 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 408-411).

[2] Xiao, L., Wang, G., & Zuo, Y. (2018). Research on patent text classification based on word2vec and LSTM. *In 2018 11th International Symposium on Computational Intelligence and Design (ISCID)* (Vol. 1, pp. 71-74).

[3] Li, S., Hu, J., Cui, Y., & Hu, J. (2018). DeepPatent: patent classification with convolutional neural networks and word embedding. *Scientometrics*, 117(2).

[4] Risch J. & Krestel, R. (2019). Domain-specific word embeddings for patent classification. *Data Technologies and Applications*

[5] Zhou, Z. H., Wu, J., & Tang, W. (2002). Ensembling neural networks: many could be better than all. *Artificial intelligence*, 137, 239-263.

[6] Mathiassen, H., & Ortiz-Arroyo, D. (2006). Automatic categorization of patent applications using classifier combinations. *In International Conference on Intelligent Data Engineering and Automated Learning* (pp. 1039-1047).

[7] Benites, F., Malmasi, S., & Zampieri, M. (2018). Classifying patent applications with ensemble methods. *arXiv preprint arXiv:1811.04695*.

[8] Kamateri, E., Stamatis, V., Diamantaras, K., & Salampasis, M. (2022). Automated Single-Label Patent Classification using Ensemble Classifiers. *ICMLC 2022.*

[9] Sofean, M. (2021). Deep learning based pipeline with multichannel inputs for patent classification. *World Patent Information*, 66, 102060.

[10] Tikk, D., Biró, G., & Törcsvári, A. (2008). A hierarchical online classifier for patent categorization. *In Emerging technologies of text mining: Techniques and applications* (pp. 244-267).

[1] https://github.com/ekamater/CLEFIP2011_XML2MySQL

Patent Classification using Extreme Multi-label Learning: A Case Study of French Patents

You Zuo
Inria Paris
Paris, France
you.zuo@inria.fr

Houda Mouzoun
Institut national de la propriété industrielle
Paris, France
hmouzoun@inpi.fr

Samir Ghamri Doudane
Institut national de la propriété industrielle
Paris, France
sghamridoudane@inpi.fr

Kim Gerdes
LISN, CNRS and University Paris-Saclay
Orsay, France
gerdes@lisn.fr

Benoît Sagot
Inria Paris
Paris, France
benoit.sagot@inria.fr

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; • **Information systems** → *Document representation*; • **Social and professional topics** → *Patents.*

KEYWORDS

IPC prediction, Clustering and classification, Extreme Multi-label Learning, French

1 INTRODUCTION

The number of patent applications has risen sharply over the past 20 years. As a result, automatic patent classification systems have become essential for patent specialists to analyze and manage large collections of patents. There are several standard classification structures, the most commonly used being the IPC (International Patent Classification) and the CPC (Cooperative Patent Classification), which have hierarchical structures with five different levels: sections, classes, subclasses, groups, and subgroups.

Most previous approaches [1, 6, 11, 12, 20–22] treat the patent classification task as a general text classification task and apply commonly used text classification methods. Some have attempted to implement XML (Extreme Multi-label Learning) methods to handle large numbers of classes [5, 24], but they focus only on the IPC subclasse level, which is far from "extreme" with less than 700 labels.

In this paper, we present a French Patents corpus, named **INPI-CLS**, with IPC labels at all levels, and we test different models at the subclass and group levels on it. Our published French patents are extracted from the INPI[1] internal database, and contain all parts of patent texts (title, abstract, claims, description) published from 2002 to 2021, each patent being annotated with all levels from sections to the IPC subgroup labels. A statistical overview of the data is given in Tables 1 and 2. The training set is constructed from patent documents published before 2020, while the test set includes patents published in 2020 and 2021. In Table 2, N represents the number of patents in the training and test sets. L indicates the label count, $\bar{L}$ stands for the average number of IPC labels of a document. $\hat{L}$ represents the average number of documents per label. The subscripts of 4,6,8 represent respectively IPC's subclass, group, and subgroup levels (4, 6, and 8 correspond to the number of characters used to encode the class). We then compare the performance of the XML (Extreme Multi-label Learning) approaches with other popular NLP methods on our INPI-CLS as well as on the English patent classification benchmark USPTO-2M[12] with 1.9 million training data and 48,000 test data.

We are releasing all relevant code and our French patent classification dataset as open source. The dataset may be used for research purposes and is available under specific licensing requirements detailed in the GitHub repository. [2]

section	title	abstract	description	claims
# items	296 270	295 421	296 216	291 539
# tokens (average)	11	111	4202	725

Table 1: Description of our French corpus INPI_fr

Dataset	N	L_4	$\bar{L_4}$	$\hat{L_4}$	L_6	$\bar{L_6}$	$\hat{L_6}$	L_8	$\bar{L_8}$	$\hat{L_8}$
Train	268254	638	1.73	420.46	6788	2.21	39.52	48932	2.73	5.48
Test	28017	583	1.77	48.06	4351	2.20	6.44	19593	2.64	1.43

Table 2: Basic Statistics of INPI-CLS dataset

2 EXPERIMENTS AND RESULTS

The details of the selected model are listed below:

Logistic Regression The one-vs-all approach is implemented to train a binary logistic regression classifier for each label. We use TF-IDF as input features after applying the snowball stemmer from NLTK and eliminating stop words from the first 1000 words of the input text.

FastText text classification [9] FastText applies a shallow neural network on a hidden variable represented by the average of n-gram character embeddings of input, where the ngram character embeddings are trained under supervision specifically for text classification. We initialize the token representations by the embedding matrix pre-trained on Wikipedia[3] and train a linear classifier for multi-label text classification.

[1] French National Institute of Industrial Property https://www.inpi.fr/fr

[2] https://github.com/ZoeYou/Patent-Classification-2022

[3] https://fasttext.cc/docs/en/pretrained-vectors.html

Published in "Proceedings of The 3rd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2022)", Eds. H. Aras, R. Krestel, F. Piroi, L. Andersson, A. Hanbury, D. Alderucci, PatentSemTech 2022, 15 July 2022, Madrid, Spain. https://doi.org/10.34726/3550 | © Authors 2022. CC BY-NC-ND 4.0 License.

Bert [4] Just as in the PatentBert[11] experiments, we fine-tune Bert on patent classification. We test the `bert-large` instead of `bert-base` to allow for a comparison with Bert for Patents [23] of the same architecture.

Bert for Patents [23] The model was trained from scratch on more than 100 million English patent documents of USPTO, it leveraged `bert-large` architecture, and built a patent-specific custom tokenizer to hold longer tokens. We took their officially released checkpoint[4] and fine-tuned it on USPTO-2M[12]. [5]

XML-CNN [13] Based on CNN-Kim[10], XML-CNN applies a dynamic maximum pooling to accommodate longer texts and extract location information. It adds a hidden bottleneck layer between the pooling and the output layer, which learns a better representation of the document and improves the prediction accuracy.[6]

Parabel [19] As one of the baseline tree-based algorithms of XML approaches, Parabel firstly learns a balanced binary tree of labels by recursively dividing the label nodes into two balanced clusters until the number of labels in each cluster is less than a given value, and then trains a probabilistic hierarchical multi-label model that generalizes hierarchical softmax to a multi-label setup.[7]

AttentionXML [26] AttentionXML compresses the binary partitioned label tree of [18] into shallower and wider tree to better handle larger label size. A bi-LSTM with multi-label attention mechanism is trained for each level of the tree with the first 500 words of raw text as input. The word representation layers are initialized by Glove[8] for English and French FastText trained on Wikipedia for French patents.

LightXML [8] LightXML applies multiple pre-trained language models. For each model, it concatenates the representations of the `[CLS]` in the last five hidden state as text representation, then trains a label recalling network to dynamically sample negative samples followed by a label ranking network to separate positive from negative labels.[9]

We tested all models on both English and French datasets, except for Bert and Bert for patent, two language models trained on the English corpus. We use ensemble approach for Parabel and AttentionXML with the number of ensemble being 3. For LightXML, we use three different encoders for ensemble. The encoders used for USPTO-2M are

- `bert-base-uncased`[4]
- `roberta-base`[14]
- `xlnet-base-cased`[25]

and

- `camembert-base`[15]
- `bert-base-multilingual-cased`[17]
- `xlm-roberta-base`[2]

for the INPI French patent corpus.

We employ the rank-based metrics Precision@K (P@k(%); k = 1, 3, 5) as evaluation metric following prior Multi-label text classification works. P@K are calculated for each test document and then averaged over all the documents. Due to space limitations, we only show the two main results that we test on the English Benchmark USPTO-2M and our new French dataset INPI-CLS (title+abstract as classifiers' input).

Table 3 demonstrates that LightXML achieves the best results on USPTO-2M, and Bert for Patents achieves comparable performance on it. Compared to the results obtained from [11, 22], we can conclude that we achieve state-of-the-art performance on USPTO-2M with LightXML. It is worth noting that Bert for patent is a large-scale language model specifically pre-trained on patent text from scratch. Bert is very time and resource intensive to train, and we may not be able to find a training corpus of the same size for non-English languages. Yet, the same performance can easily be achieved or even exceeded based on LightXML using ensemble learning with several other off-the-shelf language models including some blocks specifically designed for the XML task. This gives the possibility to obtain higher patent classification performance in languages that do not have as much patent data as English (e.g. French).

For our proposed French patent classification dataset INPI-CLS, LightXML is vastly outperforming the others on both subclass and group levels. LightXML's outstanding performance is attributed to its powerful feature extraction from multiple layers of different transformer encoders and its negative sampling approach on dynamically selecting negative labels from easy to difficult.

Model	P@1	P@3	P@5
Logistic Regression	74.63	41.66	28.82
FastText	73.89	40.55	28.02
bert-large	83.77	46.27	31.37
Bert for Patents	84.31	46.73	31.73
XML-CNN	57.00	31.22	22.08
Parabel	74.43	41.49	28.50
AttentionXML	82.49	45.15	30.82
LightXML	**84.43**	**46.81**	**31.91**

Table 3: Overall Performance on IPC subclass on USPTO-2M (title + abstract)

Model	subclass			group		
	P@1	P@3	P@5	P@1	P@3	P@5
Logistic Regression	65.87	37.63	26.02	49.12	30.32	22.06
FastText	53.76	30.64	21.31	36.21	22.32	16.35
XML-CNN	43.43	25.50	18.23	17.74	10.20	6.96
Parabel	65.13	36.87	25.32	48.93	30.61	22.28
AttentionXML	72.54	40.68	27.63	54.83	33.78	24.49
LightXML	**76.45**	**42.82**	**29.05**	**60.60**	**36.95**	**26.65**

Table 4: Overall Performance on IPC subclass and group on INPI-CLS (title + abstract)

The reasons why the same model performs better on USPTO-2M are 1) USPTO-2M has a much larger dataset for training, almost ten times larger than the French dataset, and 2) by calculating the KL-divergence of the label distributions of the training and test data, we find that the label distributions of the training and test

[4] BERT-for-patents GitHub repository.

[5] The hyperparameters for fine-tuning the two previous language models on patent classification are set as follows: max_sequence_length = 128; epoch = 4; batch_size = 32; learning_rate (Adam) = $3e^{-5}$; binary cross-entropy loss.

[6] We used the code provided by the authors with default values for hyperparameters from https://github.com/siddsax/XML-CNN.

[7] The scripts we utilize are from the Omikuji project. We change CBOW to TF-IDF for better label representation and leave all other hyperparameters as default.

[8] Glove 840B,300d from https://nlp.stanford.edu/projects/glove/

[9] For AttentionXML and LightXML, we used codes provided by the online extreme classification repository.

data are closer for USPTO-2M than that for INPI-CLS. Therefore, we assert that our dataset is "more difficult" to classify.

Different combinations of document parts were tested on our proposed French patent corpus and it was experimentally demonstrated that the combination of title and description achieves the best results (compared to abstract, claims, description and title+abstract). More precisely, when the input constraints are loose (much larger than 128 subwords), there is an improvement of about 4% to 8% on precision@1. However, for methods using pre-trained language models with max_sequence_length set to 128, the precision improvement using title+description compared to title+abstract is less than 2%.

We perform the error analysis by examining the single-label AUC and confusion matrix at $k = 1$. We conclude that weaker models perform worse in learning to classify those labels with less training examples (the AUC of the classifier corresponding to that IPC label is lower). Also, all models have a tendency to mistake "long-tail" labels for those more frequent labels.

3 ONGOING AND FUTURE WORK

Our current focus is on classifying labels with fewer patent examples by using label descriptions or correlations between labels as input information as in [3, 16] and using propensity scored metrics [7] to better evaluate the "long-tailed" labels.

REFERENCES

[1] Juho Bai, Inwook Shim, and Seog Park. 2020. MEXN: Multi-Stage Extraction Network for Patent Document Classification. *Applied Sciences* 10, 18 (2020).

[2] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).

[3] Kunal Dahiya, Ananye Agarwal, Deepak Saini, Gururaj K, Jian Jiao, Amit Singh, Sumeet Agarwal, Purushottam Kar, and Manik Varma. 2021. SiameseXML: Siamese Networks meet Extreme Classifiers with 100M Labels. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 2330–2340.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] Arousha Haghighian Roudsari, Jafar Afshar, Charles Lee, and Wookey Lee. 2020. Multi-label Patent Classification using Attention-Aware Deep Learning Model. 558–559.

[6] Jason Hepburn. 2018. Universal Language Model Fine-tuning for Patent Classification. In *Proceedings of the Australasian Language Technology Association Workshop 2018*. Dunedin, New Zealand, 93–96.

[7] Himanshu Jain, Yashoteja Prabhu, and Manik Varma. 2016. Extreme Multi-Label Loss Functions for Recommendation, Tagging, Ranking amp; Other Missing Label Applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 935–944.

[8] Ting Jiang, Deqing Wang, Leilei Sun, Huayi Yang, Zhengyang Zhao, and Fuzhen Zhuang. 2021. Lightxml: Transformer with dynamic negative sampling for high-performance extreme multi-label text classification. *arXiv preprint arXiv:2101.03305* (2021).

[9] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification. arXiv:1607.01759 [cs.CL]

[10] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification.

[11] Jieh-Sheng Lee and Jieh Hsiang. 2019. PatentBERT: Patent Classification with Fine-Tuning a pre-trained BERT Model. *CoRR* abs/1906.02124 (2019).

[12] Shaobo Li, Jie Hu, Yuxin Cui, and Jianjun Hu. 2018. DeepPatent: Patent Classification with Convolutional Neural Networks and Word Embedding. *Scientometrics* 117, 2 (nov 2018), 721–744.

[13] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep Learning for Extreme Multi-label Text Classification. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2017).

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint:1907.11692* (2019).

[15] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of ACL*.

[16] Anshul Mittal, Noveen Sachdeva, Sheshansh Agrawal, Sumeet Agarwal, Purushottam Kar, and Manik Varma. 2021. ECLARE: Extreme Classification with Label Graph Correlations. *CoRR* abs/2108.00261 (2021).

[17] Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? *arXiv preprint arXiv:1906.01502* (2019).

[18] Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. 2018. Parabel. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*. ACM Press.

[19] Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. 2018. Parabel: Partitioned Label Trees for Extreme Classification with Application to Dynamic Search Advertising. *WWW '18: Proceedings of the 2018 World Wide Web Conference*, 993–1002.

[20] Subhash Pujari, Annemarie Friedrich, and Jannik Strötgen. 2021. *A Multi-task Approach to Neural Multi-label Hierarchical Patent Classification Using Transformers*. 513–528.

[21] Julian Risch, Samuele Garda, and Ralf Krestel. 2020. Hierarchical Document Classification as a Sequence Generation Task. In *Proceedings of the Joint Conference on Digital Libraries (JCDL)*. 147–155.

[22] Arousha Haghighian Roudsari, Jafar Afshar, Wookey Lee, and Suan Lee. 2021. PatentNet: multi-label classification of patent documents using deep learning based language understanding. *Scientometrics* (2021).

[23] Rob Srebrovic and Jay Yonamine. 2020. *Leveraging the BERT algorithm for Patents with TensorFlow and BigQuery*. Technical Report. Global Patents, Google, https://services.google.com/fh/files/blogs/bert_for_patents_white_paper.pdf.

[24] Pingjie Tang, Meng Jiang, Bryan Ning Xia, Jed W Pitera, Jeffrey Welser, and Nitesh V Chawla. 2020. Multi-label patent categorization with non-local attention-based graph convolutional network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9024–9031.

[25] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:1906.08237 [cs.CL]

[26] Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2019. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. *Advances in Neural Information Processing Systems* 32 (2019).

Query Terms Suggestion

Applying semantic similarity measures to support query formulation for patent search

Matthias Wirth
Data Science
European Patent Office
Rijswijk, Netherlands
mwirth@epo.org

Alexander Klenner-Bajaja
Data Science
European Patent Office
Rijswijk, Netherlands
aklenner@epo.org

ABSTRACT

The European Patent Office (EPO) has defined its mission[1] "to deliver high-quality patents and efficient services that foster innovation, competitiveness and economic growth." In its Strategic Plan 2023 the EPO committed itself to increase the efficiency of the search procedure by providing its patent examiners with an improved set of tools to support their daily work.

There is a high correlation between identifying a set of very relevant documents in the Search phase and the quality of the decision that is taken concerning the patentability of an application. Despite all advances made in automated analysis of incoming applications, and automated retrieval of relevant documents, well formulated and constructed keyword queries against a complete prior art index are still the foundation of a successful search.

Therefore, providing tools to support query generation is a key task for Business Information Technology within the EPO. We set out to provide a framework to enable the retrieval of semantically similar words for a given query word, or concept.

By leveraging our large corpus of 120 million patent documents, we trained a set of word2vec[2] models for different technical areas in all our core-languages (EN, DE, FR), in which applications can be filed at the Office.

In our contribution to PatentSemTech 2022, we will showcase the developed solution, the technical complexity and motivate the strategical decisions taken.

We will focus on:

1. Data cleaning and processing
2. Technical challenges of our solution in an operational setting
3. Scenarios how the implementation will support examiners in their work

Cleanliness and correctness of data is a key requirement for the final models, as retrieved nearest words need to be of high quality, and not e.g. malformatted words due to OCR errors in the original document. We will demonstrate and motivate our implement-tations used to pre-process the datasets.

As words can have different meaning in different technical fields, we define models per technical domains guided by Collaborative Patent Classification[3] (CPC) allocations of the patent documents. We will present performed experiments to assess the granularity under which the best trade-off between quality of nearest word neighbours and number of independent models was observed.

In the context of our business, cross lingual information retrieval is an important aspect of the examiner's work, and an identified requirement is to look up semantically similar words in different languages. This can be achieved via alignment of the resulting word embedding space of the different models. We will showcase how this is supported with our implementation approach.

Another important requirement is the support of concepts queries: Identifying potential query terms not yet known to the human searcher for a given set of keyword queries. We will demonstrate use cases of how concept queries can fill blind spots and complement and assist human defined queries.

Finally, we will present our current strategy to support our users by an up-to-date terminology of the models. We implemented a fully automated machine learning operations pipeline that can be run in defined intervals or event triggered to ensure the availability of high quality models in our productions tools. We demonstrate how model and data lineage is implemented, automated monitoring of the training process is handled and how the model is deployed in operations. We will present the underlying architecture and the selected technologies for our setup at the EPO.

REFERENCES

[1] European Patent Office. (2022,0404). https://www.epo.org/about-us/office/mission.html
[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean (2013) Efficient Estimation of Word Representations in Vector Space, https://arxiv.org/abs/1301.3781
[3] European Patent Office. (2022,0404). https://www.epo.org/searching-for-patents/helpful-resources/first-time-here/classification/cpc.html

Published in "Proceedings of The 3rd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2022)", Eds. H. Aras, R. Krestel, F. Piroi, L. Andersson, A. Hanbury, D. Alderucci, PatentSemTech 2022, 15 July 2022, Madrid, Spain. https://doi.org/10.34726/3550 | © Authors 2022. CC BY-NC-ND 4.0 License.

A Combination of BERT and BM25 for Patent Search

Vasileios Stamatis
International Hellenic University
Thessaloniki, Greece
vstamatis@the.ihu.gr

Michail Salampasis
International Hellenic University
Thessaloniki, Greece
msa@ihu.gr

Konstantinos Diamantaras
International Hellenic University
Thessaloniki, Greece
kdiamant@ihu.gr

Allan Hanbury
Vienna University of Technology
Vienna, Austria
allan.hanbury@tuwien.ac.at

1 INTRODUCTION

Several methods for patent search have been developed in the last years [1, 2]. Researchers have been using various techniques that are query-oriented, meta-data based, Pseudo Relevance Feedback Methods, Semantic-based methods, etc. Recently there has been a shift in research towards implementing machine learning methods for the patent domain [3, 4]. However, transformer models like BERT [7] that have achieved impressive results on various NLP tasks have not been sufficiently explored for patent search. While BERT has drawn some attention in research in the patent industry, it is either used for classification [8, 9] or didn't work as much effectively as expected for retrieval [10]. Our research investigates how to adapt the BERT model in the patent domain to increase the retrieval performance.

Specifically, the research question we will explore in this work is: how can the BERT model be adapted to improve retrieval effectiveness in patent prior art search? To do this, first, we examine BERT's generalization ability to the patent domain in a zero-shot setting and demonstrate that it cannot improve patent retrieval effectiveness as an off-the-shelf method. Then, we propose a new hybrid document retrieval method and test it in patent prior art search task. The proposed technique combines BM25's and BERT's scores so that the BERT model is used as a scaling factor that operates on the BM25 score and modifies it according to the BERT estimate of relevance. To train a BERT model in patent language, we also created a new dataset that consists of relevant and non-relevant pairs of patent abstracts, called Intellectual Property Abstracts (IPA dataset). This dataset is a processed extract from the MAREC dataset [11], and is available for download[*].

CCS CONCEPTS

• Information systems → Information retrieval → Retrieval models and ranking • Computing methodologies → Artificial intelligence → Natural language processing

KEYWORDS

Prior Art Search, Patent Search, BERT, Neural IR, Domain-Specific Search

2 METHODOLOGY

The experiments presented in this paper are based on the CLEF-IP 2011 collection [12] and our new IPA dataset. To create the IPA dataset, we iterate all MAREC documents, and for each document with an English abstract, we process its citations. For every citation, we extracted it's English abstract and wrote one relevant instance in the CSV file (abstract_doc | abstract_citation | 1) and one non-relevant instance using the abstract from a random document from the MAREC collection (abstract_doc | abstract_random | 0). Finally, we removed all the lines containing abstracts used in CLEF-IP topics so that our retrieval results will not be biased. The whole dataset contains approximately 78 million pairs of abstracts.

The queries we used are taken from the 3973 topics of the CLEF-IP 2011 evaluation campaign. We used the first 150 English topics. Each query consists of a maximum of 500 words produced sequentially from the title, abstract, description, and the claims.

2.1 Hybrid Retrieval method

The proposed approach fuses the effectiveness of lexical methods and the deep semantic similarities that the neural models can capture. Fang & Zhai [13] examined the semantic term matching constraint, which states that the exact matching of a term is as much important for the relevance score as matching a semantically related term several times. Even though this work was presented some time ago, lexical models still perform better than semantic methods such as BERT in the patent domain as we observe in our experiments and the literature [10]. Our hybrid approach combines the lexical signals from BM25 as the main relevance factor and a fine-tuned cross-encoder BERT model that will score the similarity between the query and the candidate documents by looking only at their abstracts. The BERT score is then combined with the BM25 score so that BERT functions as a scaling factor, operating on the BM25 score. The final score will be an increased or decreased or with no change BM25 score. Mathematically, we get the final scores using the formula (1) below:

$$score = bm25 + c * bm25 * bert \qquad (1)$$

where $bm25$ is the BM25 score and $bert$ is the BERT relevance

[*] https://zenodo.org/record/6612936#.YptqxHZBy1s

Published in "Proceedings of The 3rd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2022)", Eds. H. Aras, R. Krestel, F. Piroi, L. Andersson, A. Hanbury, D. Alderucci, PatentSemTech 2022, 15 July 2022, Madrid, Spain. https://doi.org/10.34726/3550 | © Authors 2022. CC BY-NC-ND 4.0 License

score. We also add the weighting factor *c* in order to reduce the bias of the different scaling between the scores. For example the BM25 score was on average between (200, and 1000), and the BERT score was between (-3, and 3). For our proposed method, we optimized the parameter c in formula (1) and conducted a grid search with different values of c (0.1, 0.25, 0.5, 0.75, 1, 2). We chose these values experimentally, observing where the scores increase and decrease. Formula (1) optimized for c = 0.25 using our dataset. For instance, if BERT estimates a candidate document as non-relevant with a low value i.e. -3, then the score would be:

$$score = bm25 - 3 * 0.25 * bm25$$

which will significantly reduce the score compared to the BM25 score and more specifically the final score will be BM25/4. Respectively, if BERT estimates a high score, i.e., 3, it will increase the final score to 1.75 * BM25. This is why we consider BERT as a scaling factor that modifies the BM25 score based on its estimate of relevance. The architecture is presented below in Figure 1.

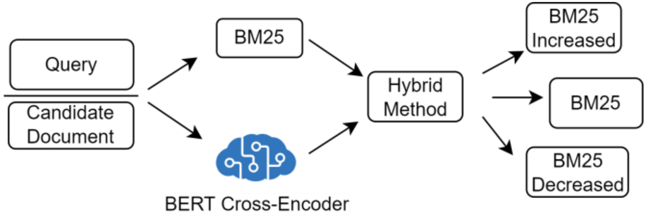


Figure 1: Hybrid Retrieval method

4 Results

To examine the effectiveness of the proposed architecture (Hybrid), we compare it with the following baselines: BM25, Cross-Encoder BERT (CE BERT), i.e., use a cross-encoder BERT model and use as inputs the abstracts between query and candidate document to get the final score. Bi-Encoder BERT (BE BERT) i.e. use the bi-encoder BERT model and use the abstract embeddings to calculate the cosine similarity between query and candidate document.

The used metrics are the MAP, RECALL, and PRES scores. As there are millions of documents in our index, we conduct an initial ranking using BM25 to retrieve 1000 documents and apply our methods to re-rank these results for efficient computation. The top-ranked 100 results are then used for comparison. Additionally, to examine BERT's ability to generalize to the patent domain in a zero-shot setting, we run the experiments twice, one with zero-shot learning settings and one with the fine-tuned BERT on the IPA dataset. We also implemented a random ranking to create the lowest baseline where the 1000 results are re-ranked randomly.

Our fine-tuned hybrid method achieved the best scores and outperformed all the baselines, especially the BM25 by 5.56% at MAP, 3.6% at PRES, and 3.5% at RECALL @100. We also conducted statistical tests and we found that even though our proposed method achieved higher scores than BM25 the results are not statistically significant.

MODEL	MAP @100	PRES @100	RECALL @100
BM25	0.0881	0.2115	0.2761
CE BERT (zero-shot)	0.0005	0.0050	0.0090
CE BERT (fine-tuned)	0.0088	0.0877	0.1544
BE BERT (fine-tuned)	0.0114	0.0521	0.0916
BE BERT (zero-shot)	0.0226	0.0868	0.1242
Hybrid (zero-shot)	0.0006	0.0045	0.0069
Hybrid (fine-tuned)	0.0930	0.2191	0.2859
Random	0.0017	0.0188	0.0421

Table 1: Results of the different models

5 Conclusion

In this work, we explored the research question of how can BERT model be adapted to improve retrieval effectiveness in patent prior art search. We adapted BERT to patent characteristics by first using only the abstracts to create a new dataset (IPA dataset) specifically for training the patent-specific BERT model. Second, we proposed a hybrid model that effectively combines BM25 and BERT models. To the best of our knowledge, this is the first time that BERT has achieved a better performance than BM25 for patent prior art search and the first time that the BERT model operates as a scaling factor for the BM25 score.

ACKNOWLEDGMENTS

We want to thank Nvidia for providing us with a Titan Xp GPU. This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No: 860721 (DoSSIER Project, https://dossier-project.eu/).

REFERENCES

[1] M. Lupu and A. Hanbury, "Patent Retrieval," Foundations and Trends in Information Retreival, vol. 7, no. 1, pp. 1-97, 2013.
[2] W. Shalaby and W. Zadrozny, "Patent retrieval: a literature review," Knowledge and Information Systems, vol. 61, pp. 631-660, 2019.
[3] D. Alderucci and D. Sicker, "Applying artificial intelligence to the patent system," Technology and Innovation, vol. 20, pp. 415-425, 2019.
[4] R. Setchi, I. Spasic, J. Morgan, C. Harrison and R. Corken, "Artificial intelligence for patent prior art searching," World Patent Information, vol. 64, 2021.
[5] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in arXiv:1810.04805v2, 2019.
[6] D. M. Kang, C. C. Lee, S. Lee and W. Lee, "Patent Prior Art Search using Deep Learning Language Model," in 24th International Database Application & Enginnerring Symosium (IDEAS 2020), ACM, New York, NY, USA, 5 pages, 2020.
[7] J.-S. Lee and H. Jieh, "Patent classification by fine-tuning BERT language model," World Patent Information, vol. 61, 2020.
[8] S. Althammer, S. Hofstatter and A. Hanbury, "Cross-domain Retrieval in the Legal and Patent Domains: a Reproducability Study," in ECIR, 2021.
[9] "MAREC data set," [Online]. Available: http://www.ifs.tuwien.ac.at/imp/marec.shtml.
[10] F. Piroi, M. Lupu, A. Hanbury and V. Zenz, "CLEF-IP 2011: Retrieval in the intellectual property," Amsterdam, The Netherlands, 2011.
[11] H. Fang and C. Zhai, "Semantic Term Matching in Axiomatic Approaches to Information Retrieval," in SIGIR, Seattle, Washington, USA, 2006.

Patent Search Using Triplet Networks Based Fine-Tuned SciBERT

Utku Umur Acikalin
u.acikalin@etu.edu.tr
TOBB University of Economics and Technology
Ankara, Turkey

Mucahid Kutlu
m.kutlu@etu.edu.tr
TOBB University of Economics and Technology
Ankara, Turkey

ABSTRACT

In this paper, we propose a novel method for the prior-art search task. We fine-tune SciBERT transformer model using Triplet Network approach, allowing us to represent each patent with a fixed-size vector. This also enables us to conduct efficient vector similarity computations to rank patents in query time. In our experiments, we show that our proposed method outperforms baseline methods.

CCS CONCEPTS

• **Information systems → Information retrieval**; • **Applied computing → Document searching**.

KEYWORDS

patent search, transformer models, information retrieval

1 INTRODUCTION

The number of patents is increasing rapidly with the incredible advances in scientific knowledge and technology. This brings many challenges for patent examiners as they have to compare each patent application against prior ones and determine whether it is novel. Therefore, we need effective search engines that can find relevant patents for a given patent application.

Prior-art search has particular challenges compared to typical search operations [5]. Firstly, the patent documents are generally long and use very technical language. Secondly, the documents are prepared to show the novelty of the application, instead of focusing on the similarities with the existing ones. Thirdly, it is a recall-oriented retrieval task as we have to find all relevant patents to detect the novelty of a patent application.

Prior work shows that BERT [2] based models achieve state-of-the-art results in various Natural Language Processing (NLP) tasks. Therefore, in order to find relevant patents for a given patent, we can fine-tune a BERT [2] model to directly predict the pairwise patent similarities. However, this approach has two main shortcomings. Firstly, BERT models are capable of processing only 512 “tokens",which corresponds to roughly 400 words on an average text. However, patents are generally much longer than 400 words and we have to provide two patent documents to calculate their similarity, reducing the number of tokens we can use for each patent. Therefore, this approach would force us to ignore many parts of patent documents. Secondly, given that we have millions of patents, predicting similarity scores using a fine-tuned BERT model for all patents for a given query patent would be excessively slow.

In this paper, we develop a novel method for overcoming the shortcomings discussed above. In particular, we represent each patent using SciBERT [1] allowing us to capture technical language used in patent documents. Next, we fine-tune SciBERT model based on Triplet Networks approach [3]. This allows us to derive a fixed vector for each patent document and apply efficient vector computations. In query time, we rank patents based on their cosine similarity to the query patent. In our experiments with 1.8M patents, we show that our proposed method outperforms baseline methods.

2 PROPOSED APPROACH

In this section, we explain the details of representing patents with SciBERT and Triplet Network based fine-tuning.

2.1 Patent Representation

BERT models are successful at catching the semantics of texts. However, the language of patent documents might include many technical terminologies while BERT is pre-trained using Wikipedia articles and BooksCorpus. Therefore, we exploit SciBERT [1], which is pre-trained on large multi-domain corpus of scientific publications, instead of using the original BERT.

Patent documents are generally much longer than BERT based models can process. We could truncate patent documents to meet the limits of BERT. However, it would mean ignoring many parts of patent documents that might be useful for our search task. Therefore, in order to capture the semantics of patent documents, we create separate embeddings for the description (v_d) and claims (v_c) part of each patent. For descriptions longer than 400 words, we use TextRank [4] automatic summarization tool to reduce the text length to 400 words. However, for the claims part of patents, we do not use the text summarization but truncate the parts that exceed BERT's token limit. This is because the first claim of patents is generally the main innovative part of the patents while the other claims are less important ones. Subsequently, we concatenate the vectors for the description and claim parts to form a single embedding for each patent and normalize them to have a unit norm. In order to give more emphasis to the description part of the patents than their claims, we multiply each element of v_d by $\sqrt{0.8}$ and multiply each element of v_c with $\sqrt{0.2}$. The parameters are selected arbitrarily. Note that because of the vector multiplication in cosine similarity calculation, the relative weights used for description and claims parts will be 8:2.

2.2 Fine-Tuning via Triplet Networks

We fine-tune SciBERT using Triplet Networks approach [3] which allows us to derive fixed-size embeddings for each patent, and thereby, apply efficient vector operations to calculate the similarity between patents. In the Triplet Network approach, we have to provide positive and negative samples for each patent such that the model can learn the semantic differences between relevant and not relevant patents. In particular, we construct 3 embeddings for each patent based on i) an anchor (i.e., the patent itself) patent (a),

Published in "Proceedings of The 3rd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2022)", Eds. H. Aras, R. Krestel, F. Piroi, L. Andersson, A. Hanbury, D. Alderucci, PatentSemTech 2022, 15 July 2022, Madrid, Spain. https://doi.org/10.34726/3550 | © Authors 2022. CC BY-NC-ND 4.0 License

Ranking Method	Average Precision	Recall@100	Recall@500	Recall@1000
Lucene with TF-IDF	0.0548	0.2178	0.3642	0.4364
Lucene with BM25	0.0469	0.1800	0.3083	0.3743
Our Approach	**0.0675**	**0.2233**	**0.3934**	**0.4821**

Table 1: Comparison of our approach with baseline methods. The best performing score for each metric is written in bold.

ii) a positive (i.e., relevant) patent (*p*), and iii) a negative (i.e., not relevant) patent (*n*). We calculate triplet objective loss as follows:

$$max(CosineDistance(v_a, v_p) - CosineDistance(v_a, v_n) + \epsilon, 0)$$

where v_a, v_p, and v_n are the embeddings for *a*, *p*, and *n*, respectively. ϵ is a margin ensuring that v_p is at least ϵ closer to v_a, than v_n.

Obviously, the training data and the label distribution directly affect supervised models' performance. Therefore, we take the following steps to select the patents given as positive and negative samples.

- We select 'positive' texts from the cited patents which have a similarity score of higher than 0.6 according to vectors provided by Google[1].
- We select 20% of the negatives from the not-cited patents which are from the Cooperative Patent Classification (CPC) group of the anchor patent. Therefore, the model can learn textual properties of patents that are on a similar topic but not as close as the positive ones.
- We select 20% of the negatives from the patents which are not cited by the anchor patent but cited by the patents that it cites. This process allows us to train the models with negative samples that are not semantically far from the anchor patent.
- The remaining 60% of the negatives are randomly selected from the patents which are not cited by the anchor patent and have a similarity score of less than 0.6 based on Google's vectors. Therefore, the model can learn the textual properties of patents that are distinctively different from the anchor.

3 EXPERIMENTS

We randomly select 2 million patents granted after 1980. Among these patents, 1,817,504 of them have a title, abstract, description, and claims sections. From this sample, we randomly select 5,000 patents for testing, and others are used in training. Following prior work [5], we consider cited patents as relevant ones and not-cited ones as not-relevant.

We train the model with four million examples (i.e., patent triplets). We use patents which have at least five backward and forward citations in total, as anchors in the training set. We train the model using 4 Nvidia Titan RTX GPUs with a batch size of 8, using Adam optimizer with a learning rate of $3e^{-6}$ with linear learning rate warm-up over 10% of the training data for 1 epoch.

We compare our model against BM25 and TF-IDF ranking functions that Lucene[2] provides. The results are shown in **Table 1**. We observe that our approach outperforms Lucene's methods based on all four metrics, suggesting that our proposed method can be an effective solution for the prior-art search problem.

4 CONCLUSION

In this paper, we propose a novel method to represent patent documents by fine-tuning SciBERT with Triplet Network approach. We show that our proposed method outperforms baseline methods in our experiments. In the future, we plan to extend our work in several directions. Firstly, we plan to use other variants of BERT pre-trained with different types of documents, e.g., PatentBERT. In addition, we plan to investigate which parts of patent documents are more important for the prior-art search task and how to best summarize them. Furthermore, we will investigate using BERT variants that have higher token limits. Finally, we believe that our model should be evaluated in various test collections and compared against other baseline methods.

ACKNOWLEDGMENT

This study was funded by the Scientific and Technological Research Council of Turkey (TUBITAK) ARDEB 1001 Grant No 119K986. The statements made herein are solely the responsibility of the authors.

REFERENCES

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3613–3618. https://doi.org/10.18653/v1/D19-1371

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[3] Elad Hoffer and Nir Ailon. 2015. Deep Metric Learning Using Triplet Network. In *Similarity-Based Pattern Recognition - Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015, Proceedings (Lecture Notes in Computer Science, Vol. 9370)*, Aasa Feragen, Marcello Pelillo, and Marco Loog (Eds.). Springer, 84–92. https://doi.org/10.1007/978-3-319-24261-3_7

[4] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain.* ACL, 404–411. https://aclanthology.org/W04-3252/

[5] Xiaobing Xue and W Bruce Croft. 2009. Automatic query generation for patent search. In *Proceedings of the 18th ACM conference on Information and knowledge management.* 2037–2040.

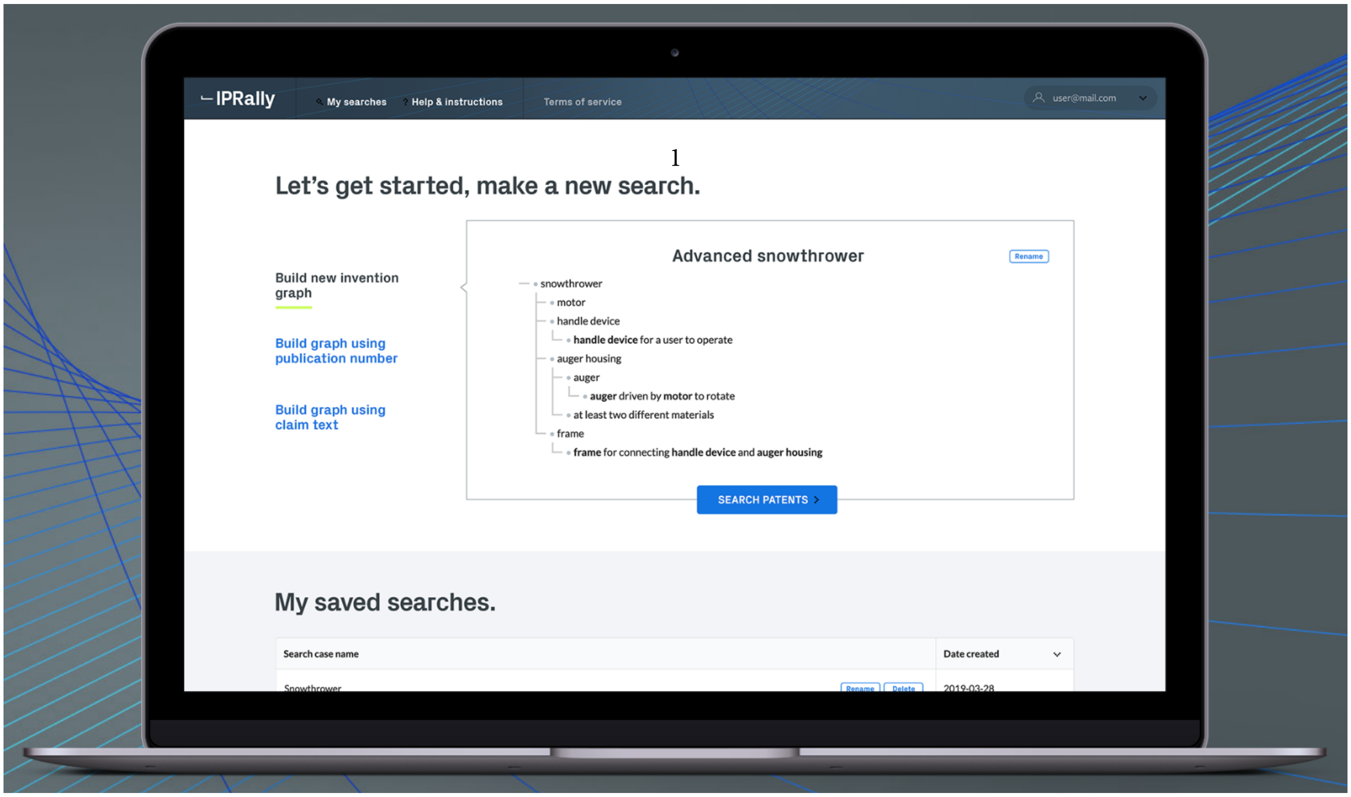
[1] https://console.cloud.google.com/marketplace/details/google_patents_public_datasets/google-patents-research-data

[2] https://lucene.apache.org/core/

Graph-based patent search

Juho Kallio
juho@iprally.com
IPRally Technologies Oy
Helsinki, Finland

Sebastian Björkqvist
sebastian@iprally.com
IPRally Technologies Oy
Helsinki, Finland



Let's get started, make a new search.

Advanced snowthrower

- snowthrower
  - motor
  - handle device
    - **handle device** for a user to operate
  - auger housing
    - auger
      - **auger** driven by **motor** to rotate
    - at least two different materials
  - frame
    - **frame** for connecting **handle device** and **auger housing**

My saved searches.

Search case name	Date created
Snowthrower	2019-03-28

Figure 1. IPRally patent search platform.

Abstract

Patent data is a unique data set for information retrieval. The data is publicly available, patent examiners from patent offices have created a network of citations and also classification for the documents, and the language in the documents is standardised to describe technology in an exact way in the legal sense.

IPRally patent search platform was developed to solve the challenge of information retrieval from patents. Our key insight is the graph format that can represent an invention in an compressed way. Only the essential features are kept from the original text, semantically connected to each others. We created a parser that converts patent document into a high quality graph - the parsing of a full patent document takes on average 10 seconds. In a cloud computing environment we are able to scale this for the 100 million patent documents that are now covered by IPRally search. We use IFI CLAIMS as our data provider.

For the general approach for the patent search, we use a vector search model [1]. The nearest neighbors of the embedding of query graph can be fetched in milliseconds using approximated nearest neighbor methods. We chose the Spotify's Annoy algorithm [2] for this.

We utilise patent examiner citations for supervised training of the encoder network. By maximizing cosine similarity of the patent application and the cited publication we can

Published in "Proceedings of The 3rd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2022)", Eds. H. Aras, R. Krestel, F. Piroi, L. Andersson, A. Hanbury, D. Alderucci, PatentSemTech 2022, 15 July 2022, Madrid, Spain. https://doi.org/10.34726/3550 | © Authors 2022. CC BY-NC-ND 4.0 License.

create a neural network that mimics a human patent examiner.

The graphs are simplified into trees. This way, TreeLSTM [4] is a suitable model architecture for the encoder network. Node values are text, which is presented with GloVe [3] word embeddings.

We measure search performance mainly with top 50 recall, considering the examiner citations. The motivation for this metric is that the typical user is expected to read similar number of results. As long as the relevant result is within top 50, the user gains the needed knowledge.

The result quality is significantly better than baseline with bag-of-words type approach. Result quality has been also the main selling point of the product - it compares favorably with the competition.

Keywords: Patent search, knowledge graph, GNN

Acknowledgement

The idea of the knowledge graph format originates from Sakari Arvela's previous career as patent attorney. IPRally was founded in 2018 by Sakari, Juho Kallio and Juuso Piskonen, and the early team also included Sebastian Björkqvist and Juho Sepänmaa. We are ever thankful for the courageous early investors: Icebreaker VC, five angel investors, and the public funding from the Finnish government.

References

[1] David Dubin. 2004. The most influential paper Gerard Salton never wrote. *Library Trends* (2004), 2004. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.184.910

[2] Wen Li, Ying Zhang, Yifang Sun, Wei Wang, Wenjie Zhang, and Xuemin Lin. 2016. Approximate Nearest Neighbor Search on High Dimensional Data — Experiments, Analyses, and Improvement (v1.0). https://doi.org/10.48550/ARXIV.1610.02455

[3] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation.. In *EMNLP*, Vol. 14. 1532–1543.

[4] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. https://doi.org/10.48550/ARXIV.1503.00075

Patents Phrase to Phrase Semantic Matching Dataset

Grigor Aslanyan
aslanyan@google.com
Google Inc.

Ian Wetherbee
wetherbeei@google.com
Google Inc.

ABSTRACT

There are many general purpose benchmark datasets for Semantic Textual Similarity but none of them are focused on technical concepts found in patents and scientific publications. This work aims to fill this gap by presenting a new human rated contextual phrase to phrase matching dataset. The entire dataset contains close to 50, 000 rated phrase pairs, each with a CPC (Cooperative Patent Classification) class as a context. This paper describes the dataset and some baseline models.

ACM Reference Format:
Grigor Aslanyan and Ian Wetherbee. 2022. Patents Phrase to Phrase Semantic Matching Dataset. In *Proceedings of SIGIR PatentSemTech Workshop.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION AND RELATED WORK

Semantic Textual Similarity (STS) measures how similar two pieces of text are. STS is one of the most important tasks in Natural Language Processing (NLP) and there has been a significant amount of research in recent years in this domain. Benchmark datasets play the important role of allowing to consistently and fairly measure model improvements. There are multiple benchmark datasets for STS that are commonly used to measure the performance of state of the art models. Some notable examples are STS-B [1], SICK [5], MRPC [3], and PIT [13]. However, these datasets are fairly general purpuse, and to the best of our knowledge there are currently no datasets that are focused on technical concepts found in patents and scientific publications. The somewhat related BioASQ challenge contains a biomedical question answering task [12].

This paper introduces a new human rated contextual phrase to phrase matching dataset focused on technical terms from patents. In addition to similarity scores that are typically included in other benchmark datasets we include granular rating classes similar to WordNet [8], such as synonym, antonym, hypernym, hyponym, holonym, meronym, domain related.

The dataset was generated with focus on the following:

- Phrase disambiguation: certain keywords and phrases can have multiple different meanings. For example, the phrase "mouse" may refer to an animal or a computer input device. We have included a context CPC class that can help disambiguate the anchor and target phrase.
- Keyword match: there are phrases that have matching keywords but are otherwise unrelated (e.g. “container section” → “kitchen container”, “offset table” → “table fan”). Many models will not do well on such data (e.g. bag of words models). Our dataset is designed to include many such examples.
- State of the art language models: We created our dataset with the aim to improve upon current state of the art language models. Specifically, we have used the BERT model [2] to generate target phrases. So our dataset contains many human rated examples of phrase pairs that BERT may identify as very similar but in fact they may not be.

anchor	target	context	rating	score
acid absorption	absorption of acid	B08	exact	1.00
acid absorption	acid immersion	B08	synonym	0.75
acid absorption	chemically soaked	B08	domain	0.25
acid absorption	acid reflux	B08	not rel.	0.00
gasoline blend	petrol blend	C10	synonym	0.75
gasoline blend	fuel blend	C10	hypernym	0.50
gasoline blend	fruit blend	C10	not rel.	0.00
faucet assembly	water tap	A22	hyponym	0.50
faucet assembly	water supply	A22	holonym	0.25
faucet assembly	school assembly	A22	not rel.	0.00

Table 1: A small sample of the data.

The dataset is used in the *U.S. Patent Phrase to Phrase Matching* Kaggle competition[1] from March 21 - June 20, 2022. After the completion of the competition the full dataset will be made public.

2 DATASET DESCRIPTION

Each entry of the dataset contains two phrases - anchor and target, a context CPC class, a rating class, and a similarity score. A small sample of the dataset is shown in Table 1.

The entire dataset contains 48, 548 entries with 973 unique anchors, split into a training (75%), validation (5%), and test (20%) sets. When splitting the data all of the entries with the same anchor are kept together in the same set. There are 106 different context CPC classes and all of them are represented in the training set.

We have used the following steps for generating the data. For each patent in the corpus we first extract important (salient) phrases. These are typically noun phrases (e.g. “fastener”, “lifting assembly”) or functional phrases (e.g. “food processing”, “ink printing”). Next, we keep only phrases that appear in at least 100 patents. We randomly sample around 1,000 phrases from the remaining phrases which become our anchor phrases. For each anchor phrase we find all of the matching patents and all of the CPC classes for those patents. From all of the matching CPC classes we randomly sample up to four. These become the context CPC classes for that anchor phrase. The target phrases come from two sources - pre-generated and rater generated.

Published in "Proceedings of The 3rd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2022)", Eds. H. Aras, R. Krestel, F. Piroi, L. Andersson, A. Hanbury, D. Alderucci, PatentSemTech 2022, 15 July 2022, Madrid, Spain. https://doi.org/10.34726/3550 | © Authors 2022. CC BY-NC-ND 4.0 License.

[1] https://www.kaggle.com/c/us-patent-phrase-to-phrase-matching

Model	Dim.	Pearson cor.	Spearman cor.
GloVe	300	0.429	0.444
FastText	300	0.402	0.467
Word2Vec	250	0.437	0.483
BERT	1024	0.418	0.409
Patent-BERT	1024	0.528	0.535
Sentence-BERT	768	0.598	0.577

Table 2: Baseline model metrics.

We use two different methods for pre-generating target phrases - partial matching and a masked language model (MLM). For partial match we randomly select phrases from the entire corpus that partially match with the anchor phrase. This means that one or more of the tokens matches, but the whole phrase is different (e.g. “abatement” → “noise abatement”, “material formation” → “formation material”). For MLM we select sentences from the patents that contain a given anchor phrase, mask it out, and use a BERT model [2] to predict candidates for the masked portion of the text. All of the phrases are cleaned up before sending to the raters. This includes lowercasing and removal of punctuation and certain stopwords (e.g. "and", "or", "said").

The raters were asked to determine the similarity level between the two phrases given the context CPC class. They choose between five different levels of similarity - very high, high, medium, low, and not related. Each similarity level is further divided into different subclasses, such as hyponym (broad-narrow), hypernym (narrow-broad), antonym, domain related. The detailed description of the similarity levels and subclasses will be included with the public release of the data.

All of the pre-generated target phrases were independently rated by two raters. After completing the ratings they met and went over all of the non-matching ratings to discuss and agree on a final rating. Each rater separately generated new target phrases and gave ratings to them. We have merged all of the rater generated phrases together. We have left out the rare cases where the two raters generated the same target phrase with different ratings.

3 BASELINES

Table 2 describe the performance of some common off the shelf models on the test data. We have only included dual-tower model architectures that perform an embedding of the anchor and target phrases separately and compute similarity using cosine distance. All of the models use mean pooling of individual keyword embeddings to get the full phrase embedding.

For GloVe [9] we have used the *Wikipedia 2014 + Gigaword 5* model[2], for FastText [4] the *wiki-news-300d-1M* model[3] [7], and for Word2Vec [6] the *Wiki-words-250* model from TensorFlow Hub[4]. For BERT [2] we have used the BERT-Large model from TensorFlow Hub[5]. For comparison, we have also included the publicly available BERT model pre-trained on patent data [11] of the same size as BERT-Large. Finally, for Sentence-BERT [10] we have used the *all-mpnet-base-v2* pretrained model[6].

The bag-of-words models do not perform very well, which is expected given the dataset structure (e.g. many matching terms with different meanings). The Patent-BERT model significantly outperforms the regular BERT model, which implies that generic pretrained models are not optimal for technical terms found in patents. However, we get the best results from the Sentence-BERT model. This is not entirely surprising since Sentence-BERT has been specifically fine tuned for the dual-tower architecture we are using for similarity.

REFERENCES

[1] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 1–14. https://doi.org/10.18653/v1/S17-2001

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[3] William B. Dolan and Chris Brockett. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*. https://aclanthology.org/I05-5002

[4] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, 427–431. https://aclanthology.org/E17-2068

[5] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland, 216–223. http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv e-prints*, Article arXiv:1301.3781 (Jan. 2013), arXiv:1301.3781 pages. arXiv:cs.CL/1301.3781

[7] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

[8] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (nov 1995), 39–41. https://doi.org/10.1145/219717.219748

[9] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[10] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. http://arxiv.org/abs/1908.10084

[11] Rob Srebrovic and Jay Yonamine. [n.d.]. Leveraging the BERT algorithm for Patents with TensorFlow and BigQuery. ([n. d.]). https://services.google.com/fh/files/blogs/bert_for_patents_white_paper.pdf

[12] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel Ngonga, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics* 16 (2015), 138. https://doi.org/10.1186/s12859-015-0564-6

[13] Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015. SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter (PIT). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, Denver, Colorado, 1–11. https://doi.org/10.18653/v1/S15-2001

[2] https://nlp.stanford.edu/projects/glove/

[3] https://fasttext.cc/docs/en/english-vectors.html

[4] https://tfhub.dev/google/Wiki-words-250/2

[5] https://tfhub.dev/tensorflow/bert_en_uncased_L-24_H-1024_A-16/4

[6] https://www.sbert.net/docs/pretrained_models.html

Recent Developments in AI and USPTO Open Data

Scott Beliveau*
Scott.Beliveau@uspto.gov
United States Patent and Trademark Office
Alexandria, Virginia, USA

Jerry Ma*
Jerry.Ma@uspto.gov
United States Patent and Trademark Office
Alexandria, Virginia, USA

ABSTRACT

The USPTO disseminates one of the largest publicly accessible repositories of scientific, technical, and commercial data worldwide. USPTO data has historically seen frequent use in fields such as patent analytics, economics, and prosecution & litigation tools. This article highlights an emerging class of usecases directed to the research, development, and application of artificial intelligence technology. Such usecases contemplate both the delivery of artificial intelligence capabilities for practical IP applications and the enablement of future state-of-the-art artificial intelligence research via USPTO data products. Examples from both within and beyond the USPTO are offered as case studies.

ACM Reference Format:
Scott Beliveau and Jerry Ma. 2022. Recent Developments in AI and USPTO Open Data. In *Proceedings of SIGIR 3rd Workshop on Patent Text Mining and Semantic Technologies (SIGIR '22 PatentSemTech).* ACM, Madrid, Spain, 3 pages.

1 INTRODUCTION

As America's national IP office, the United States Patent and Trademark Office (USPTO) is charged with the mission of granting patents and registering trademarks. The USPTO is required by law to disseminate most nonprovisional patent applications and granted patents to the public. [1] U.S. patent data has long been used in many domains of application, including in patent analytics [18], economics [2, 4], and commercial tools for patent prosecutors and litigators, [2] thus serving as a versatile substrate for illustrating the dynamics of national and global innovation.

Concurrently, the fields of artificial intelligence (AI) and natural language processing (NLP) have witnessed a remarkable cadence of scientific breakthroughs. Fueled by model architecture innovations such as self-attention [26] and by the ever-increasing computational horsepower of the leading AI hardware accelerators [21, 27], these novel techniques have found versatile areas of application, from machine translation [26] to structural biology [20].

In this article, we survey recent developments at the intersection of AI and USPTO data. These developments fall in two broad categories:

(1) Promising AI and NLP techniques can be brought to bear on USPTO data in existing or novel fields of application.
(2) USPTO data can contribute to work that advances the frontiers of AI and NLP research.

Both bodies of work hold great promise for advancing scientific and technical progress. We encourage those in both the AI and the IP communities to explore how USPTO data can unlock numerous exciting opportunities—both in their respective disciplines and at the intersection of AI & IP.

2 USPTO DATA FOR PATENT-FOCUSED AI & NLP USECASES

One distinct body of work uses AI & NLP techniques on USPTO patent data toward enhancing the value of such data toward longstanding areas of application. Here, we present case studies in the context of IP administration, practice, and empirical research.

2.1 AI & NLP tools for IP administration

IP offices worldwide seek to apply AI & NLP in the administration of their respective IP systems, and the USPTO is no exception. The USPTO recognizes the advent of AI as among the most consequential technologies—both for global society as a whole and for the agency's mission of delivering reliable, timely, and quality IP rights [13].

Operationally, the USPTO focuses on two critical areas of AI application: prior art search and patent classification. AI is a natural tool with which to augment prior art search systems. Representation learning and related techniques can produce semantically meaningful embeddings of language, graphs, images, and even proteins [8–10, 20]. The USPTO applies such techniques on the agency's patent archives and uses the results toward improving examiner-facing search systems to surface more relevant prior art documents [24].

Turning to patent classification, the USPTO currently classifies patents using a two-stage process. First, the agency assigns a set of Cooperative Patent Classification (CPC) symbols to each patent to characterize the relevant technologies contained therein. Second, the agency determines the subset of CPC symbols ("claim indicators") associated with claim scope. The USPTO has recently deployed an AI system, trained on annotated USPTO patent data, for assigning claim indicators, and the agency is currently augmenting the system to assign the full set of CPC symbols [13].

[*] The views expressed in this extended abstract should not be construed as official policy statements of the United States Patent and Trademark Office or of the U.S. Government. All errors are the authors' own.
[1] 35 U.S.C. § 122(b); 37 C.F.R. § 1.11(a).
[2] Because agency practice is to refrain from public endorsements of any particular commercial IP products, we omit references to specific examples.

Published in "Proceedings of The 3rd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2022)", Eds. H. Aras, R. Krestel, F. Piroi, L. Andersson, A. Hanbury, D. Alderucci, PatentSemTech 2022, 15 July 2022, Madrid, Spain. https://doi.org/10.34726/3550 | © Authors 2022. CC BY-NC-ND 4.0 License.

2.2 AI & NLP tools for IP practitioners & inventors

Since the dawn of computer-based information retrieval, software developers have built tools for assisting IP practitioners and inventors. Some tools are similar to those needed by IP offices (*e.g.*, prior art search), while others are specific to the needs of the private IP bar and inventors (*e.g.*, IP portfolio intelligence). Recent work has used publicly-available USPTO data to train AI models that provide new or enhanced capabilities to IP software products.

The USPTO has recently released AI-empowered search capabilities to the public through the Inventor Search Assistant [25]. This tool surfaces not only published applications from the USPTO patent archives, but also non-patent literature (NPL) and foreign patent documents. Traditional prior art search systems have a steep learning curve (*e.g.*, basic query syntax, proximity operators) that may pose a hurdle to early-stage and independent inventors. Such inventors can especially benefit from the Inventor Search Assistant, which uses machine learning techniques to offer an initial overview of the state-of-the-art from natural language queries alone.

2.3 AI-powered empirical research & analytics

Finally, USPTO data can be elucidated via existing AI & NLP techniques to produce boundary-pushing empirical research & analytics. A common patent analysis task is to sort patent documents into specific fields of technology or business applications—commonly known as "patent landscaping" [23]. Recent work has applied deep learning to the task of patent landscaping [1], with USPTO data frequently used both as training data and as the source of documents to be landscaped. The USPTO has recently leveraged such techniques in its own empirical studies on U.S. patent archives.

Released in 2021, the USPTO's AI Patent Dataset identifies the presence of AI in over 13 million U.S. patent documents and further subcategorizes them into one of eight component technologies [12]. This dataset was created by training a recurrent neural network in a semi-supervised manner to distinguish between positive and negative examples [1]. The USPTO leveraged the AI Patent Dataset to trace the diffusion of AI and its component technologies within post-1976 U.S. patents [22], with such findings informing agency stakeholder engagements and other policy-relevant activities [24].

Much patent analysis focuses on specifications, claims, and metadata, but an often-overlooked data source for patent analytics lies in prosecution history. The USPTO has applied AI techniques to make Office actions more accessible to the patent analysis community. Released in 2017, the USPTO's Office Action Research Dataset comprises a relational database of key elements from 4.4 million Office actions mailed during the 2008 to mid-2017 period [17]. This dataset was created using machine learning and NLP techniques to systematically extract information from Office actions, thus marking the first time that comprehensive data on examiner-issued rejections was made readily available to the research community.

3 ADVANCING THE AI & NLP RESEARCH FRONTIERS VIA USPTO DATA

The foregoing body of work centers around the application of existing AI & NLP techniques in IP-relevant areas. But another emerging body of work flips this paradigm by using USPTO data as an accelerant for scientific research in AI & NLP. We highlight examples in both the training and evaluation of AI models.

3.1 USPTO data in large language modeling

Large language models have demonstrated a surprisingly diverse portfolio of natural language capabilities [7, 8]. Yet early iterations of billion-parameter language models employed lightly curated datasets constructed with few quality or diversity filters. Observing this, Gao et al. [11] compiled a dataset prioritizing both data quality and diversity, combining the background sections of millions of U.S. patents with 21 other data sources to form "The Pile".

This 825 GiB language modeling dataset, and subsets thereof, were subsequently used in training or assessing some of the largest and most advanced language models to appear in published research, including GPT-NeoX-20B, Gopher, and RETRO [5, 6, 19]. OPT-175B [29], currently the largest publicly-available language model by parameter count, was trained on public USPTO data sourced from The Pile.

We observe that the background sections of patents, while informative, only scratch the surface of available content within the U.S. patent archives. Future language modeling datasets could include full patent specifications or prosecution history documents (*e.g.*, Office actions). The latter holds particular promise as a source of scientific and legal reasoning examples not easily found elsewhere.

3.2 Patent-sourced datasets for common tasks

The quantity and detail of patent documents also readily enable their use as datasets for common AI & NLP benchmark tasks. Patent classification (discussed in Section 2.1) is, at its core, a quintessential multiclass classification challenge. AI researchers have already used public USPTO data and CPC annotations to create text classification benchmarks encompassing millions of patent documents [15, 16]. These benchmarks have subsequently been used to evaluate the capabilities of new self-attention neural network models [28].

Recent work has also augmented public USPTO data with automated data generation and manual annotations to form specialized benchmark datasets that can test the ability of novel AI and NLP models to penetrate complex technical concepts. For instance, Aslanyan and Wetherbee [3] construct a novel semantic similarity benchmark dataset by extrating phrases from patent documents, generating facially similar phrases, and manually rating the semantic similarity of each phrase pair on a five-point scale. A Kaggle competition featuring this benchmark resulted in nearly 43,000 submissions, achieving a top Pearson correlation of 87.8% [14]. The USPTO is interested in building upon these early successes by fostering future efforts to refashion patent data into valuable AI research benchmarks.

4 CONCLUSION

We have described two technical bodies of work that rest upon USPTO data. The first integrates USPTO data with AI & NLP techniques to benefit IP administration, practice, and empirical analysis. The second leverages USPTO data in service of state-of-the-art AI & NLP research.

We envision these two spheres forming a virtuous cycle wherein successes in one area furthers progress in the other. From search engines to benchmarks, and from landscapes to large language models and beyond, we hope that researchers and practitioners will find novel means of harnessing the richness of USPTO data to serve both the IP community and future AI researchers.

ACKNOWLEDGMENTS

We acknowledge our stellar colleagues from across the USPTO who have labored tirelessly to advance the agency's data dissemination and artificial intelligence efforts.

REFERENCES

[1] Aaron Abood and Dave Feltenberger. 2018. Automated patent landscaping. *Artificial Intelligence and Law* 26, 2 (2018), 103–125.

[2] Ufuk Akcigit, John Grigsby, Tom Nicholas, and Stefanie Stantcheva. 2018. *Taxation and Innovation in the 20th Century.* Technical Report. National Bureau of Economic Research.

[3] Grigor Aslanyan and Ian Wetherbee. 2022. Patents Phrase to Phrase Semantic Matching Dataset. In *SIGIR 3rd Workshop on Patent Text Mining and Semantic Technologies*.

[4] Alex Bell, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen. 2018. Who Becomes an Inventor in America? The Importance of Exposure to Innovation*. *The Quarterly Journal of Economics* 134, 2 (11 2018), 647–713. https://doi.org/10.1093/qje/qjy028

[5] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An Open-Source Autoregressive Language Model. *CoRR* abs/2204.06745 (2022). https://doi.org/10.48550/arXiv.2204.06745 arXiv:2204.06745

[6] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2021. Improving language models by retrieving from trillions of tokens. *CoRR* abs/2112.04426 (2021). arXiv:2112.04426 https://arxiv.org/abs/2112.04426

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.).

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[10] Vijay Prakash Dwivedi and Xavier Bresson. 2021. A Generalization of Transformer Networks to Graphs. *AAAI Workshop on Deep Learning on Graphs: Methods and Applications* (2021).

[11] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv preprint arXiv:2101.00027* (2020).

[12] Alexander V Giczy, Nicholas A Pairolero, and Andrew A Toole. 2022. Identifying artificial intelligence (AI) invention: A novel AI patent dataset. *The Journal of Technology Transfer* 47, 2 (2022), 476–505.

[13] Drew Hirshfeld. 2021. Artificial intelligence tools at the USPTO. *Director's Blog: the latest from the USPTO leadership* (March 2021).

[14] Kaggle. 2022. U.S. Patent Phrase to Phrase Matching.

[15] Jieh-Sheng Lee and Jieh Hsiang. 2020. Patent classification by fine-tuning BERT language model. *World Patent Information* 61 (2020), 101965. https://doi.org/10.1016/j.wpi.2020.101965

[16] Shaobo Li, Jie Hu, Yuxin Cui, and Jianjun Hu. 2018. DeepPatent: patent classification with convolutional neural networks and word embedding. *Scientometrics* 117, 2 (2018), 721–744.

[17] Qiang Lu, Amanda Myers, and Scott Beliveau. 2017. USPTO patent prosecution research data: Unlocking office action traits. (2017).

[18] Paul Oldham et al. 2016. *WIPO Manual on Open Source Tools for Patent Analytics.* Technical Report.

[19] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *CoRR* abs/2112.11446 (2021). arXiv:2112.11446

[20] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* 118, 15 (2021), e2016239118. https://doi.org/10.1073/pnas.2016239118

[21] Martin Svedin, Steven Wei Der Chien, Gibson Chikafa, Niclas Jansson, and Artur Podobas. 2021. Benchmarking the Nvidia GPU Lineage. *CoRR* abs/2106.04979 (2021). arXiv:2106.04979

[22] Andrew Toole, Nicholas Pairolero, Alexander Giczy, James Forman, Christyann Pulliam, Matthew Such, Kakali Chaki, David Orange, Anne Thomas Homescu, Jesse Frumkin, Ying Yu Chen, Vincent Gonzales, Christian Hannon, Steve Melnick, Eric Nilsson, and Ben Rifkin. 2020. Inventing AI: Tracing the diffusion of artificial intelligence with U.S. patents.

[23] Anthony Trippe. 2015. Guidelines for preparing patent landscape reports. *Patent landscape reports. Geneva: WIPO* (2015), 2015.

[24] U.S. Patent and Trademark Office. 2022. Inaugural Meeting of the AI and Emerging Technologies Partnership Series.

[25] U.S. Patent and Trademark Office. 2022. Start Your Search With the Inventor Search Assistant Tool. *Inventors' Digest* (January 2022), 6.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008.

[27] Yu Wang, Gu-Yeon Wei, and David Brooks. 2019. Benchmarking TPU, GPU, and CPU Platforms for Deep Learning. *CoRR* abs/1907.10701 (2019). arXiv:1907.10701

[28] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for Longer Sequences. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.).

[29] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. *CoRR* abs/2205.01068 (2022). https://doi.org/10.48550/arXiv.2205.01068 arXiv:2205.01068