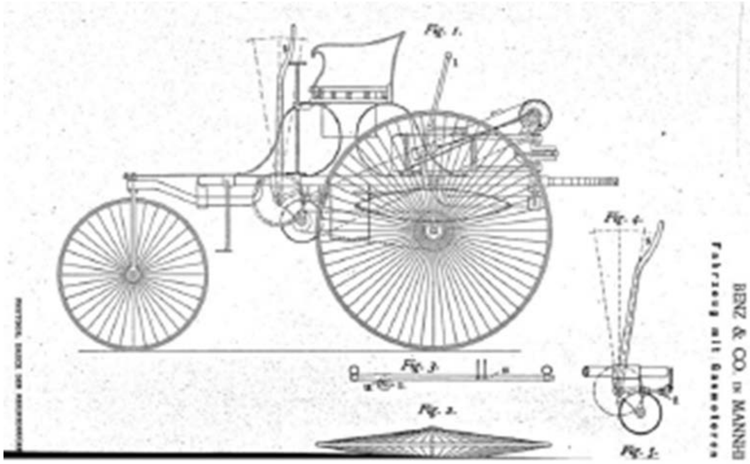
Proceedings of the

1st Workshop on on Patent Text Mining and Semantic Technologies

Editors: Linda Andersson
Hidir Aras
Florina Piroi
Allan Hanbury



PatentSemTech2019

Karlsruhe, Germany
September 12th, 2019

Editors:

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

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

Florina Piroi, Florina.Piroi@tuwien.ac.at, Florina.Piroi@artificialresearcher.com
Institute of Information Systems Engineering, Technische Universität Wien,
Vienna, Austria
Artificial Researcher IT GmbH, Vienna, Austria

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

This document contains the proceedings of the 1st Workshop on on Patent Text Mining and Semantic Technologies (PatentSemTech 2019) helod on September 12, 2019 in Karlsruhe, Germany. All submissions to this workshop have gone through single-blind reviewing process.

The workshop was organized as a one day event and had the endorsement from the European Patent Office. The workshop was co-located with the SEMANTiCS 2019 conference that took place in September 09-12, 2019, in Karslruhe, Germany.

The editors would like to thank all the scientific committe members of PatentSemTech2019.

DOI: 10.34726/pst2019

Contents

1st Workshop on Patent Text Mining and Semantic Technologies (**PatentSemTech 2019**)

Hidir Aras[1]

FIZ Karlsruhe – Leibniz Institute for Information Infrastructure

Linda Andersson, Florina Piroi[2]

TU Wien, Institute for Information Systems Engineering

Abstract

This volume presents the proceedings of the 1st Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2019) co-located with the SEMANTiCS 2019 conference, held in Karlsruhe, Germany. It is a first in series of workshops that aims to establish a long-term collaboration and a two-way communication channel between the IP industry and academia from relevant fields to foster the usage of semantic technologies for answering research questions related to patent text mining and patent analytics as well as adopt them in working applications.

Keywords: Patent Text Mining, Semantic Technologies, Semantics 2019, Intellectual Property, Machine Learning

1. Introduction

The PatentSemTech 2019 workshop[3] is a first in series of workshops that aims to establish a long-term collaboration and a two-way communication channel between the IP industry and academia from relevant fields such as natural-language processing (NLP), text and data mining (TDM) and semantic technologies (ST) in order to explore and transfer new knowledge, methods and technologies for the benefit of industrial applications as well as support research in applied sciences for the IP and neighbouring domains.

We invited scientific contributions as well as proof of concepts that show relevant use cases for patent text mining and analytics. Moreover, we invited researchers to investigate and promote new means for bootstrapping training data generation, e.g. for labelling domain-specific data sets from the Intellectual Property (IP) domain. The articles included in this volume went through a peer-review process where each submission was reviewed by at least three reviewers out of a mixed programme committee of academic researchers and experts from the IP domain. Seven papers passed the review process – 3 long, 2 short and 2 demo papers. Three submissions that passed the reviewing process were proposed for publication to the World Patent Information[4] (WPI) virtual special issue on "Patent Text Mining and Semantic Technologies". For these submissions we only included their (extended) abstracts in these proceedings.

In its first year, the workshop was organized as a one day event. We have invited as a keynote speaker Mr. A. Trippe, managing director of Patinformatics LLC., a long year, internationally recognised IP expert, and an adjunct Professor of IP Management and Markets at Illinois Institute of Technology where he teaches courses on patent analysis and landscapes for strategic decision making. His keynote addressed the importance of patent analytics tools based on semantics and machine learning techniques for the strategic decisions that businesses need to take with respect to their long term

[1] https://www.fiz-karlsruhe.de

[2] https://www.ifs.tuwien.ac.at

[3] http://www.ifs.tuwien.ac.at/patentsemtech/

[4] https://www.journals.elsevier.com/world-patent-information

Published in "Proceedings of The 1st Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2019)", Eds. L. Andersson, H. Aras, F. Piroi, A. Hanbury, PatentSemTech 2019, 12 September 2019, Karlsruhe, Germany. https://doi.org/10.34726/pst2019.1 | © Authors 2019. CC BY 4.0 License.

R&D and economic plans.

2. Keynote: Improving Patent Analytics Using Semantic Technologies

The use of patent analytics has increased exponentially over the past ten years. So much so that even the worlds patent offices have devoted resources and staff to create departments responsible for developing insights into technology areas of importance to that country or region using output generated applying patent analytics. At the same time new tools, methods and systems have begun to emerge that seek to make the analysis of patent data easier to accomplish. Included in these new developments are a significant number of approaches that apply machine learning and make use of knowledge modeling and semantic analysis in order to deal with existing challenges for text and data analytics. As these changes continue to occur, it would be useful to review a list of the tasks associated with patent analytics and think about the types of tools, systems, or methods that a patent analyst would like to have at their disposal. Starting with a general overview of patent analytics, and with a focus on patent landscape reports, case studies and perspectives will be provided on why this work is so highly valued. The presentation concluded with a prioritized list of suggestions for how patent analytics and patent landscape creation could be aided by the further development and implementation of semantic technologies.

3. Main Topics and Objectives

In the started workshop series we aim to set the basis for researchers and the IP industry to explore next-generation text and data mining methods and semantic technologies for the enrichment and large-scale analysis of huge amounts (Big Data) of scientific-technical information in general and patent data in particular.

We want to motivate and enable scientists from academia to make use of and exploit the richness of the scientific-technical information that is amassed nowhere else but in the patent data, by, for example, interlinking it to other knowledge sources from domain-specific knowledge graphs (bio-pharma, chemistry or engineering, etc.) or the linked open data cloud.

Starting with publicly available datasets for patent mining and patent retrieval tasks such as classification, passage retrieval, etc. we want to set the focus on developing enhanced methods for analysing patent texts by applying machine learning and making use of implicit and explicit semantic information.

Hence, the workshop series aims to motivate research and development in related areas in order to

- *explore* IP applications with underlying advanced NLP, TDM and artificial intelligence methods, e.g. applying Deep Learning (DL) for generating patent embeddings, etc.
- *apply* enhanced machine or deep learning technologies for the semantic enrichment and analysis of big data of patent texts, e.g. to contribute to use cases such as technology analysis, trend analysis, semantic patent landscaping, competitor analysis, etc.
- *show* proof of concepts for patent and technology analysis use cases such as patent landscaping, portfolio analysis, white and hotspot analysis, technology trends analysis, etc.
- *evaluate* new visual user interface concepts for exploring and analysing large datasets of scientific texts

There have been several text mining initiatives in terms of establishing tools and benchmark collections for widely used data such as news corpora, medical data, etc. However, a set of benchmark collections covering the diversity of the information needs of the IP industry, as for example detailed in [30, 4], is still missing. A long term goal of this series of workshops is therefore to encourage future research collaboration with focus on IP related data – patent documents, non-patent literature, court litigation cases – and combine it with more traditional patent analytic resources, like meta-data, to be used for the above described tasks and use cases.

4. State of the Art and the Impact of Training and Test Data

Patent text mining [48, 24, 22, 36] includes research on the handling and the integration of multiple and diverse information sources, since information related to IP for science and technology are siloed into various repositories consisting of laws, regulations, patents, court litigation, scientific publications etc.

4.1. Patent Retrieval

As a research field, Patent Retrieval belongs to domain-specific information retrieval, hence, represents a sub discipline of information retrieval (IR). The research focus of patent retrieval is to develop techniques and methods that effectively and efficiently retrieve relevant patent documents or paragraphs in response to an information need [36, 50, 51]. IR has received a significant amount of focus from researchers in different computer science disciplines since many decades. In comparison, patent retrieval is poorly treated by the academic scientific communities, with periodic surges of such activities whenever patent data became available to researchers. It is only during the last 20 years that the challenges in patent retrieval have been a target for the research community [30]. On reason for this is that, compared to other types of text, the patent genre presents unique features such as lengthy documents (multi-page), multi-modal documents (e.g. image, text, algorithm and programming codes), multi-language, semi-structured, meta-data rich, stretching over a variety of technologies (heterogeneous). Answers to information needs in IP also vary from a complete multi-page application to a one-page inventor disclosure to just a few keywords [22].

In a scientific context patent retrieval was first introduced in the NIIs NTCIR-1 campaigns (2002 to 2007) [15], followed by several initiatives that included patent retrieval as a research topic, e.g. Dutch Belgian Information Retrieval workshop [41], ASPIRE [17], Patent Information Retrieval (PaIR), TREC-Chem (TExt REtrieval Conference Chemical track) [28], and the Information Retrieval Facility Symposium and Conference (2008-2014) [42, 29]. The largest academic research impact, in Europe, has been made by the CLEF-IP track, which was part of the Cross-Language Evaluation Forum (CLEF). The CLEF-IP track was organized from 2009 to 2013 and included a variety of tasks ranging from image classification, prior art search, and patent text classification.

4.2. Passage Retrieval

In 2012, CLEF-IP introduced the Patent Passage Retrieval task [33]. Given a patent application and selected claims in the document, the aim was to retrieve relevant documents *and* also extract those paragraphs (passages) from them that are found most relevant. Since CLEF-IP used mainly data released by the European Patent Office (EPO), the relevance assessments were semi-automatically extracted from EPO search reports.

The passage retrieval task is very close in spirit with the work of patent examiners done during an invalidity or validity search: examiners need to identify both the prior art documents, as well as each specific paragraph within these documents considered to be Prior Art for specific claims in the patent application [18, 36]. Patent Passage Retrieval could be seen as a cross-over between adhoc document retrieval tasks and question answering (QA) tasks. Concretely, in order to achieve good performance it is required that query formulations include automatic technical term extraction, followed by an advanced ad-hoc IR approach. Furthermore, in order to narrow in on each relevant passage information extraction (IE) approaches needs to be considered as well.

4.3. Enhanced Semantic Analysis and Patent Mining

Segmenting the full text of patent documents (e.g. patent descriptions [39] text, claims [16]) is regarded an important step for the semantic structure analysis of the patent texts. New approaches based on machine learning are increasingly used for a variety of tasks related to patent text mining and large-scale patent analytics [38]. Important examples are trends analysis [47], technology forecasting [43], various clustering algorithms like reinforcement learning [10], support vector clustering [49], and matrix factorization [13].

In the last few years researchers started to apply Deep Learning methods to patent text mining tasks such as keyword extraction [21], synonym extraction [25] or patent classification [12]. Tasks such as calculating patent similarity [37], patent segmentation [11], and patent landscaping [3] can be considered as important sub-tasks to be considered. In addition, various types of embedding [32] such as graph embedding [46], word embedding have been applied to evaluating patent similarity [9] or text classification tasks [44].

4.4. Benchmark Data and Patent Resources for Patent Mining Tasks

After the CLEF-IP and the TREC-Chem evaluation campaigns, and the test collections resulting out of them, further efforts to establish and update benchmark text collections have been made, the latest is the WPI collection [26].

In the World Patent Information (WPI) journal, own data collections are provided in order to support the objectives of the journal, to publish new research and insights covering a broad spectrum of intellectual property information retrieval and patent analytics related practices and methods. The WPI journal editors together with the team at IFI CLAIMS Patent Services have put together a patent research collection, publicly available and for free, to foster scientific good practice: comparability, reproducibility, transparency and repeatability of experiments and results.

The WPI collection is for this reason static. It will not be updated with new data. This decision was made in order to make sure that experimental results are traceable and due to improvements of the proposed mining/retrieval methods are due to algorithmic improvements and not due to changes in the dataset. The WPI collection complements existing test collections, which are vertical (one domain or one authority over many years). Compared to them, the WPI collection is horizontal: it includes all technical domains from the major patenting authorities over the relatively short time span of two years.

Other, industry driven initiatives to establish resources for patent text mining also aim to provide researchers with benchmark data for this area:

- In October 2017, Google launched several patent related data collections and services. Google provides the Google Patents Public Datasets[5] on BigQuery, with a collection of publicly accessible, connected database tables for empirical analysis of the international patent system. The Google Patent Datasets can provide a solution to developing and answering search oriented questions. For instance, it is possible to formulate questions such as ”what percentage of the patents have more than one inventor?” or ”what funding does the government provide to promote innovation in certain patent areas?”

- Linked Open EP data[6] uses Uniform Resource Identifiers (URIs) to identify patent applications, publications and other resources present in patent data. The URIs make it possible to link the data other datasets. The data set covers the most relevant, but not all available bibliographic data elements for patents and not all data elements from the CPC scheme. It also includes references to the full text publication in PDF, HTML and XML format, which are stored on the European Publication Server. Linked Open EP data creates a public web of interlinked patent data from EPO and other data publishers that can be queried, retrieved and viewed using standardized web technologies like HTTP, URI, RDF and SPARQL.

A common problem in machine learning and particularly in the patent domain is the creation of labelled data (i.e. training and testing data) for a variety of search and analysis tasks. As available labelled corpora are either too small or not accessible to the research community, we want to alleviate this situation with a three-year effort. That is, we plan to target the creation of more datasets open to research and addressing different patent text mining and patent text analysis applications with focus on Information Extraction (IE), classification, clustering, and to establish further datasets that require only semi-supervised training methods.

Currently, the publicly available patent mining datasets involve only a few different types of IR applications (classification, passage retrieval and prior art search[7]). In order to explore and support other patent text mining and text analysis applications that originate from the diversity of IP experts' information needs, we aim for the creation of more IE-oriented datasets. The IE-oriented datasets will be designed for domain-specific terminology extraction, for example extraction of particular token types like mathematical formula, chemical compounds, quantity entity, sequences programming codes.

These datasets will provide to the industry and the research community a variety of benchmark data, a kind of 'PatentPedia', which can support different types of question answering systems ranging from text-based to knowledge-based approaches. Furthermore, a variety of patent retrieval and analysis tasks such as technology analysis, trend analysis, semantic patent landscaping,

[5] https://cloud.google.com/blog/products/gcp/google-patents-public-datasets-connecting-public-paid-and-private-patent-data

[6] https://www.epo.org/searching-for-patents/data/linked-open-data.html

[7] For example see `http://ifs.tuwien.ac.at/patentsemtech/data-sources.html`

competitor analysis, etc. could be explored, developed and evaluated.

The effort to establish these datasets will be undertaken as a community effort with an annotation task. As organisers we intend to provide a small starting set, which is gradually improved and increased as the task is launched and running. A similar procedure has been explored in BioNLP and SemEval [19] successfully.

During the first workshop we used existing data, which, though small in size, allowed us to apply supervised and unsupervised methods to detect technical terms [14]. For the coming years we plan define tasks that build on previous ones, with the aim of creating specific patent text sets of data where completing specific tasks is needed in order to approach the next one.

5. Impact and Expectations

5.1. Target Audience

The workshop is customised for the IP industry experts as well as for academic researchers. To attract the IP industry and especially expert users, we have been in contact with members of the CEPIUG (Confederacy of European Patent Information User Groups) in order to offer Continuing Professional Development (CPD) points.

For securing visibility and increase in research submissions and participation from both industry and academia, the authors of a selected set of accepted papers are invited to submit extended versions of their research to the virtual special issue of WPI, "Text Mining and Semantic Technologies in the Intellectual Property Domain". This year the following papers have been promoted to the virtual special issue:

- *Deep Learning based Pipeline with Multichannel Inputs for Patent Classification*
- *Detecting Multi Word Terms in Patents the same way as Named Entities*
- *Semantic Views - Interactive Hierarchical Exploration for Patent Landscaping*

In the future we aim to span over different scientific disciplines such as Economic[45, 23] and Social Science [20], which also have a long tradition of working with patent analytics, in particular on citation networks. Furthermore, we would like to involve, the Triz community *Invention Innovation creativity and design*, which has run their own conference since 2009. We are working to engage national and international patent offices, such as the EPO, and companies active in the patent industry (e.g. Legit, Patsnap, Landscaping Valuenex, Google Patent, to name a few) and invite them to the workshop events.

Our key speaker in the first workshop in the series we invited a representative of the IP industry side. In our second, upcoming event, we plan to invite an academic key speaker, while in the third event we plan for a panel debate with participants from patent offices, industry and academics.

Mr. Anthony Trippe, the key speaker of the first event, is Managing Director of Patinformatics, LLC. Patinformatics is an advisory firm specializing in patent analytics and landscaping to support decision making for technology based businesses. In addition to operating Patinformatics, Mr. Trippe is also an Adjunct Professor of IP Management and Markets at Illinois Institute of Technology teaching a course on patent analysis, and landscapes for strategic decision making. He has written or contributed to IP related articles that have appeared in the Wall Street Journal, Forbes, The Washington Post, and more than a dozen additional sources. Besides that, Mr Trippe has worked for a variety of organizations, in a number of different capacities, including: P&G where he was responsible for evangelizing the use of patent analytics for business decision, Aurigin Systems where he travelled the world working with companies of all sizes that use patent analytics for competitive advantage.

5.2. How do we want to engage the patent expert to evaluate and assess our tools?

One of the key issues when developing domain-specific mining tools, especially tools requiring labelled data and expert assessment, is to engage a sufficient number of domain-experts to establish a sizable corpora or benchmark collections. Therefore, for the participating patent experts, upon request, we will issue a certificate statement which describes and documents the tasks they have been involved in, certificate signed by the head of the organisation committee. Within the certification body of the International Standard Board for Qualified Patent Information Professionals, the certified professional needs to engage in Continued Professional Development (CPD) on an annual basis. There are four types of group activities:

1. Presenting at a conference and co-author a paper on patent-related topics,
2. Participating in courses related to patent information or patentable subject matter,
3. Reading publication on patent information,
4. Peer reviewing manuscripts or search reports, or attending patent information vendor, webinar.

Within the PatentSemtTech workshop series there is ample opportunity to obtain credit for each of the group activities. Our workshop allows the IP professionals to participate as a reviewer, a presenter, or to learn about new emerging technology as well as design future use cases and contributed to establishing new benchmark collections within the field of patent text mining.

6. Organizing and Programme Committee

6.1. Organizing Committee

The organizing committee consists of persons with experience both in academic research and in close collaboration with experts in the IP domain. Two of the committee members have been key persons in organizing and running the CLEF-IP and TREC-Chem campaigns.

- Dr. Hidir Aras, FIZ Karlsruhe, Germany
- Linda Andersson, TU Wien & Artificial Researcher IT, Austria
- Dr. Lei Zhang, FIZ Karlsruhe, Germany
- Dr. Florina Piroi, TU Wien & Artificial Researcher IT, Austria
- Prof. Dr. Allan Hanbury, TU Wien, Austria
- Dr. Mihai Lupu, Data Science Studio, Research Studios Austria Forschungsgesellschaft, Austria

6.2. Programme Committee

We are grateful to the following people for providing high quality reviews and helping the workshop organizers with the submission selection process:

- Jian Wang, University of Leiden, Netherlands
- Simone Ponzetto, University of Mannheim, Germany
- Hans-Peter Zorn, inovex Gmbh, Karlsruhe, Germany
- Catherine Faron Zucker, University of Nice, France
- Ron Daniel, Elsevier Labs, USA
- Natasa Varytimou, Refinitiv, formerly Thomson Reuters, UK
- Paul Groth, University of Amsterdam, Netherlands
- Natterer Michael, Dennemeyer Octimine GmbH, Munich, Germany
- Pedro Szekeli, USC Viterbi School of Engineering, USA
- Kobkaew Opasjumruskit, German Aerospace Center, Jena, Germany
- Shariq Bashir, University of Islamabad, Pakistan
- Michail Salampasis, International Hellenic University, Greece
- Siegfried Handschuh, University of St. Gallen, Switzerland
- Agata Filipowska, Poznan University of Economics, Poland
- Rene Hackl-Sommer, FIZ Karlsruhe, Germany
- Richard Eckart de Castilho, TU Darmstadt, Germany
- Joni Sayeler, Uppdragshuset Sverige AB, Sweden
- Mustafa Sofean, FIZ Karlsruhe, Germany
- Christoph Hewel, Patent Attorney at Cabinet Beau de Lomnie, France
- Sebastian Pado, University of Stuttgart, Germany
- Parvaz Mahdabi, Swisscom, Switzerland
- Gabriela Ferraro, Australian National University, Australia
- Wlodek Zadrozny, UNC Charlotte, USA
- Bharathi Raja Chakravarthi, Insight Centre for Data Analytics, National University of Ireland, Galway

7. Conclusions

The workshop organized this year addressed researchers from academics as well as industrial experts from relevant domains and aimed to establish a two-way communication channel between both. The general feedback was very positive and participants recommended keeping the good mix of scientific and practical presentations and the demos.

The participating experts expressed that such an event was missing since a while and efforts towards this direction are welcome by both - IP experts as well as academic researchers who supported the workshop actively by, for example, providing data or participating in paper reviewing as part of the programme committee.

The PatentSemTech workshop can be seen as a first initiative to establish a patent data mining community and will be more than a one-day event per year. Our intention is to make it into an active community with webinars on relevant topics, training and assessment activities to promote patent data mining and creating benchmark data to address different patent use cases and tasks.

We plan to run the workshop for three years, and a selected set of peer-reviewed and accepted scientific papers will be invited to be published in a Virtual Special Issue (VSI) of the World Patent Information "Text Mining and Semantic Technologies in the Intellectual Property Domain". Submissions to the VSI is possible also during the years after the workshop has taken place. In addition, it is planned to establish social media channels (Twitter, LinkedIn) in order to publish news related to the workshop and future activities. We recommend all interested researchers to have a look at our workshop website, where we will update datasets and resources, or announce interesting results and events.

References

[1] Assad Abbas, Limin Zhang, Samee U. Khan, A literature review on the state-of-the-art in patent analysis, World Patent Information, Volume 37, 2014.

[2] Aras, Hidir, Ren Hackl-Sommer, Michael Schwantner and Mustafa Sofean. Applications and Challenges of Text Mining with Patents. IPaMin@KONVENS, 2014.

[3] Aaron Abood and Dave Feltenberger. 2018. Automated patent landscaping. Artif. Intell. Law 26, 2 (June 2018), 103-125.

[4] D. Alberts, C. Barcelon Yang, D. Fobare-DePonio, K. Koubek, S. Robins, M. Rodgers, E. Simmons, D. DeMarco. 2017. Introduction to Patent Searching: Practical Experience and Requirements for Searching the Patent Spaces. In [27].

[5] L. Andersson, M. Lupu, J. Palotti, A. Hanbury, and A. Rauber. When is the time ripe for natural language processing for passage patent retrieval monitoring of vocabulary shifts over time. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM 16, 2016.

[6] L. Andersson, M. Lupu, Joo R. M. Palotti, F. Piroi, A. Hanbury, and A. Rauber. Insight to hyponymy lexical relation extraction in the patent genre versus other text genres. In Proceedings of the First International Workshop on Patent Mining and Its Applications (IPaMin 2014) co-located with Konvens 2014, Hildesheim,Germany, October 6-7, 2014., 2014.

[7] Anick, P. G., M. Verhagen, and J. Pustejovsky. "Identification of Technology Terms in Patents." LREC. 2014.

[8] J. Alex, H. Schtze, and S. Brgmann. "Unsupervised training set generation for automatic acquisition of technical terminology in patents." Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical Papers. 2014.

[9] Aras, H.; Trker, R.; Geiss, D.; Milbradt, M.; Sack, H. Get Your Hands Dirty: Evaluating Word2Vec Models for Patent Data, In Proc. of the 14th Int. Conf. on Semantic Systems (SEMANTICS 2018), P&D Track, CEUR workshop proceedings vol. 2198, 2018.

[10] H. Beltz, A. Fueloep, R. R. Wadhwa, P. Erdi, From ranking and clustering of evolving networks to patent citation analysis, in: Neural Networks 350 (IJCNN), 2017 International Joint Conference on, IEEE.

[11] Carvalho, Danilo & Nguyen, Minh-Le. (2017). Efficient Neural-based patent document segmentation with Term Order Probabilities. ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017.

[12] Don, S & Min, Dugki. (2016). Feature Selection for Automatic Categorization of Patent Documents. Indian Journal of Science and Technology.

[13] R. Du, B. Drake, H. Park, Hybrid Clustering based on Content and Connection Structure using Joint Nonnegative Matrix Factorization, 2017, arXiv:1703.09646.

[14] T. Fink. Improving Multi Word Term Detection in the Patent Domain with Deep Learning. 2018 Master thesis. TU Wien

[15] Fujii, M. Iwayama, and N. Kando. Introduction to the special issue on patent processing. Information Processing & Management, 43(5):1149–1153, 2007. Patent Processing.

[16] Hackl-Sommer, Rene; Schwantner, Michael. Patent Claim Structure Recognition. Archives of Data Science, Series A, 2017, v. 2(1), 15

[17] Hanbury, V. Zenz, and H. Berger. 1st international workshop on advances in patent information retrieval (aspire10). SIGIR Forum, 44(1):1922, August 2010.

[18] D. Hunt, L. Nguyen, and M. Rodgers. Patent Searching: Tools & Techniques. Wiley, 2007.

[19] N. Ide and J. Pustejovsky, eds. Handbook of Linguistic Annotation. Springer, 2017.

[20] A. B. Jaffe, S. R. Peterson, P. R. Portney and R. N. Stavins. 1995. Environmental Regulation and the Competitiveness of U.S. Manufacturing: What Does the Evidence Tell Us? In Journal of Economic Literature. Vol. (33). No (1). pages 132-163. American Economic Association

[21] Hu, Jie, Shaobo Li, Yong Yao, Liya Yu, Guanci Yang and Jianjun Hu. Patent Keyword Extraction Algorithm Based on Distributed Representation for Patent Classification. Entropy 20 (2018): 104. Sunghae Jun, Sang-Sung Park, and Dong-Sik Jang. 2014. Document clustering method using dimension reduction and support vector clustering to overcome sparseness. Expert Syst. Appl. 41, 7 (June 2014), 3204-3212.

[22] J. Jrgens, C. Womser-Hacker, and T. Mandl. Modeling the interactive patent retrieval process: an adaptation of marchioninis information seeking model. In Fifth Information Interaction in Context Symposium, IIiX 14, Regensburg, Germany, August 26-29, 2014, pages 247250, New York, NY, USA, 2014.

[23] D. Franz, Kogler, G. Heimeriks and L. Leydesdorff. 2018 Patent portfolio analysis of cities: statistics and

maps of technological inventiveness, In European Planning Studies, Routledge, Vol. (26), No (11), pages 2256-2278, Routledge

[24] R. Krestel and P. Smyth. Recommending patents based on latent topics. In Proceedings of the 7th ACM Conference on Recommender Systems, RecSys 13, pages 395398, New York, NY, USA, 2013. ACM.

[25] A. Leeuwenberg, M. Vela, J. Dehdari, J. van Genabith. A Minimally Supervised Approach for Synonym Extraction with Word Embeddings. The Prague Bulletin of Mathematical Linguistics volume 105, Pages 111-142, De Gruyter, Berlin, Germany, 4/2016.

[26] M. Lupu, L. Papariello, R. Alentorn, M. Baycroft, J. List, The WPI patent test collection, World Patent Information, Volume 56, 2019, Pages 78-85, ISSN 0172-2190.

[27] Mihai Lupu, Katja Mayer, Noriko Kando, and Anthony J. Trippe. Current Challenges in Patent Information Retrieval (2nd ed.). Springer Publishing Company, Incorporated, 2017.

[28] M. Lupu, K. Mayer, J. Tait, and A. J. Trippe. Current Challenges in Patent Information Retrieval. Springer Publishing Company, Incorporated, 1st edition, 2011.

[29] M. Lupu, A. Hanbury, and A. Rauber. 4th international workshop on patent information retrieval (pair11). In Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011, pages 26232624.

[30] M. Lupu and A. Hanbury (2013), "Patent Retrieval", Foundations and Trends in Information Retrieval: Vol. 7: No. 1, pp 1-97.

[31] M. Lupu, J. Huang, and J. Zhu. Evaluation of chemical information retrieval tools. In Current Challenges in Patent Information Retrieval. Springer, 2011.

[32] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA, 2013.

[33] F. Piroi, M. Lupu, and A. Hanbury. Passage retrieval starting from patent claims. A clef-ip 2013 task overview. In Working Notes for CLEF 2013 Conference, Valencia, Spain, September 23-26, 2013.

[34] O. Nekhayenko. Eigennamenerkennung fr Technologien. Implementierung und Evaluierung eines Prototyps fr Patente, 2016 Master Thesis, Stiftung Universitt Hildesheim.

[35] H. Menkge. Computer Aided Patent Processing: Natural Language Processing, Machine Learning, and Information Retrieval. PhD Thesis Drexel University library 2018.

[36] W. Shalaby and W. Zadrozny. Patent retrieval: A literature review. CoRR, abs/1701.00324, 2017.

[37] P. Sharma, R. Tripathi and R. C. Tripathi, "Finding similar patents through semantic expansion," 2016 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, 2016, pp. 1-5.

[38] Sofean M., Aras H., Alrifai A. (2018) A Workflow-Based Large-Scale Patent Mining and Analytics Framework. Information and Software Technologies, In: Damaeviius R., Vasiljevien G. (eds) Information and Software Technologies. ICIST 2018. Communications in Computer and Information Science, vol 920. Springer, Cham.

[39] Sofean, M. Automatic Segmentation of Big Data of Patent Texts in: Proceedings of the International Conference on Big Data Analytics and Knowledge Discovery. DaWaK 2017. Springer pp 343-351.

[40] S. Taduri, Gloria T. Lau, Kincho H. Law, and Jay P. Kesan. A patent system ontology for facilitating retrieval of patent related information. In Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance, ICEGOV, 2012.

[41] J. Tait, M. Lupu, H. Berger, G. Roda, M. Dittenbach, A. Pesenhofer, E. Graf, and van Rijsbergen K. Patent search: An important new test bed for ir. In The Dutch-Belgian Information Retrieval Workshop, 2009.

[42] J. Tait. Information retrieval facility symposium in Vienna. SIGIR Forum, 42(1):67, 2008.

[43] C. V. Trappey, H.Y. Wu, F. Taghaboni-Dutta, and A. J. C. Trappey, Using patent data for technology forecasting: China RFID patent analysis, Advanced Engineering Informatics, 2011.

[44] Trker, R.; Zhang, L.; Koutraki, M.; Sack H. The Less Is More for Text Classification", In Proc. of the 14th Int. Conf. on Semantic Systems (SEMANTICS 2018) P&D Track, CEUR workshop proceedings vol. 2198.

[45] B. Van Looy, J. Callaert and K. Debackere. 2006. Publication and patent behavior of academic researchers: Conflicting, reinforcing or merely co-existing?. Research Policy. Vol. (35), No (4),pages 596 - 608,

[46] Wu, T.; Zhang, D.; Zhang, L.; Qi, G. Cross-Lingual Taxonomy Alignment with Bilingual Knowledge Graph Embeddings, 7th Joint International Semantic Technology Conference (JIST 2017), Cold Coast, QLD, Australia, November 10-12, 2017.

[47] J. Yoon and K. Kim, TrendPerceptor. A property function based technology intelligence system for identifying technology trends from patents, Expert Systems with Applications, 2012.

[48] H. Yu, S. Taduri, J. Kesan, G. Lau, and H. Law Kincho. Retrieving information across multiple, related domains based on user query and feedback: Application to patent laws and regulations. In Proceedings of the 4th International Conference on Theory and Practice of Electronic Governance, ICEGOV 10, pages 143151, New York, NY, USA, 2010.

[49] S. Jun, S.-S. Park, D.-S. Jang, Document clustering method using dimension reduction and support vector clustering to overcome sparseness, Expert Systems with Applications, 2014.

[50] L. Zhang, L. Li, and T. Li. Patent mining: A survey. SIGKDD Explor. Newsl. 16(2):119, May 2015.

[51] L. Zhang. An Integrated Framework for Patent Analysis and Mining. PhD thesis, Florida International University, USA, 2016.

Text Mining to Measure Novelty and Diffusion of Technological Inventions

Sam Arts* Jianan Hou** Juan Carlos Gomez***

**Department of Management, Strategy and Innovation, Faculty of Economics and Business, KU Leuven,
Korte Nieuwstraat 33, 2000 Antwerp, Belgium (e-mail: sam.arts@kuleuven.be).
** Department of Management, Strategy and Innovation, Faculty of Economics and Business, KU Leuven,
Korte Nieuwstraat 33, 2000 Antwerp, Belgium (e-mail: jianan.hou@kuleuven.be).
*** Department of Electronics Engineering, University of Guanajuato Campus Irapuato-Salamanca, Carretera Salamanca - Valle de
Santiago, Salamanca, Mexico (e-mail: jc.gomez@ugto.mx)*

Abstract: Traditional measures of patent novelty and diffusion mostly rely on patent classification or citation information. Given that inventive ideas are embedded in the text of patent documents, our study validates different alternative natural language processing techniques to measure the novelty and diffusion of technological inventions. As a validation test, we collect a set of patents linked to famous awards such as the Nobel prize. Overall, text-based measures outperform other commonly-used novelty and diffusion metrics.

Keywords: patent, text, natural language processing, novelty, impact, indicator

1. INTRODUCTION

The increasing number of granted patents echoes the prosperity of innovation activities. Nevertheless, the distribution of patent quality is highly skewed as most inventions are categorized as small incremental advances to existing technologies with little impact on subsequent invention and economic growth (Nelson and Winter, 1982; Henderson and Clark, 1990). To assess the novelty and diffusion of patents, prior studies mainly rely on patent classification or citation information (e.g. Trajtenberg 1990; Fleming, 2001; Dahlin and Behrens, 2005). The validity of traditional measures have been questioned by recent studies (McNamee, 2013; Arts et al., 2018; Kuhn and Thompson, 2019), and one of the most obvious limitations is that neither patent classifications nor citations can perfectly mirror the technological content of the patent.

In this paper, we focus on the technological content of patents and develop new patent novelty and diffusion measures by means of natural language processing techniques. To do so, we collect US patents granted up to 2018 and identify the first occurrence of a new word or word combination to pinpoint the origin of new technologies. The reuse frequency of new words and word combinations are used as indicators of technology diffusion. To examine the validity of the new measures, we collect a sample of patents awarded by prestigious prizes, such as Nobel Prize and A.M. Turing Award.

2. IDENTIFYING THE ORIGIN AND DIFFUSION OF NEW TECHNOLOGIES

We collect titles, abstracts, and claims of US utility patents granted between 1969 and 2018 from the USPTO, the patent claims research dataset (Marco et al., 2016), and PATSTAT. For all patents, we concatenate the titles, abstracts, and claims, lowercase the text, tokenize words, and remove punctuation, words composed of numbers only, one-digit words, words which appear in only one patent, and natural stop words. Then, we stem the remaining keywords and remove duplicate stemmed keywords from the same patent. Finally, the technical content of each patent is summarized by a collection of unique keywords.

Based on the processed unique words list, we trace the origin of new technologies by identifying the first patent introducing a given word or word combination. All patents are sorted by filing date, and keywords from patents filed before 1980 are used to compile the baseline dictionary (Balsmeier et al., 2018). To assess the diffusion of new technology, we count the number of subsequent US patents reusing the given new word or word combination. Finally, for each patent, we calculate the total number of new words and new word combinations as indicator of novelty and aggregate reuse frequency of all new words and new word combinations as indicators of diffusion.

*Published in "Proceedings of The 1**st Workshop on Patent Text Mining and Semantic Technologies** (**PatentSemTech 2019**)", Eds. L. Andersson, H. Aras, F. Piroi, A. Hanbury, PatentSemTech 2019, 12 September 2019, Karlsruhe, Germany. https://doi.org/10.34726/pst2019.2 | © Authors 2019. CC BY 4.0 License.*

We calculate several commonly used novelty and diffusion measures and compare their performance with the new text-based measures. First, we calculate new subclass combinations as the number of previously uncombined pairs of patent subclasses and weight it by the total number of subsequent patents reusing the focal new subclass combinations to generate a measure of diffusion (Fleming et al., 2007, Arts & Veugelers, 2014). Similarly, we count the number of previously uncombined pairs of cited patents and count the number of future patents citing the same two patents (Arts and Fleming, 2018). By examining the technological diversity of cited and citing patents, we calculate originality as one minus the Herfindahl index based on classes of cited patents, and generality as one minus the Herfindahl index based on classes of citing patents (Trajtenberg et al., 1997). Finally, we count forward citations as the number of citations received by the focal patent within 10 years (Trajtenberg, 1990).

3. VALIDATION

To assess the ability of the new text-based measures, we collect a set of patents with arguably high novelty and diffusion from seven prestigious prizes (Carpenter et al., 1981; Arts et al, 2013), namely Nobel Prize, Lasker Award, A.M. Turing Award, National Inventor Hall of Fame, National Medal of Technology and Innovation, Benjamin Franklin Medal, and Bower Award. Given that most awards (except National Inventors Hall of Fame) do not provide the patent numbers of awarded inventions, we manually match each awarded invention to US patents by the name of laureate, technical description of the awarded invention, year of discovery and laureate's affiliation. For each awarded patent, we select one control patent based on text similarity and approximate filing date (Arts et al., 2018). First, we run t tests to compare the means of the different measures for the award and control patents. Award patents score significantly higher on all measures, except for originality and new citation combination. New word combinations show the strongest discriminating power. Subsequently, we run logit regressions to predict the likelihood of being an award patent. New word combinations strongly dominates other measures in distinguishing awarded patents from control patents.

4. CONCLUSION

We develop new text mining techniques to measure the novelty and diffusion of technological inventions in the population of U.S patents. Whereas prior studies predominantly rely on patent classification or citations, we focus on the technical content of patents. By a validation test, we show that text-based measures outperform traditional measures. We will provide open access to all code and data for all US utility patents granted before May 2018.

REFERENCES

Arts, S., Appio, F. P., & Van Looy, B. (2013). Inventions shaping technological trajectories: do existing patent indicators provide a comprehensive picture? *Scientometrics*, 97(2), 397-419.

Arts, S., & Veugelers, R. (2014). Technology familiarity, recombinant novelty, and breakthrough invention. *Industrial and Corporate Change*, 24(6), 1215-1246.

Arts, S., Cassiman, B. & Gomez, J. C., (2018). Text matching to measure patent similarity. *Strategic Management Journal*, 39(1), 62-84.

Arts, S. & Fleming, L. (2018). Paradise of novelty-or loss of human capital? Exploring new fields and inventive output, *Organization Science*, 29(6), 989-1236.

Balsmeier, B., Assaf, M., Chesebro, T., Fierro, G., Johnson, K., Johnson, S., Li, G., Luck, S., O'Reagan, D., Yeh, B., Zang, G. & Fleming, L. (2018). Machine learning and natural language processing on the patent corpus: data, tools, and new measures. *Journal of Economics & Management Strategy*, 27(3), 535-553.

Carpenter, M.P., Narin, F. & Woolf, P. (1982). Citation rates to technologically important patents, *World Patent Information*, 3(4), 160-163.

Dahlin, K. B., Behrens. D. M., (2005), When is an invention really radical? defining and measuring technological radicalness, *Research Policy*, 34, 717-737.

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science*, 47(1), 117-132.

Fleming, L., Mingo, S. & Chen, D. (2007). Collaborative brokerage, generative creativity, and creative success, *Administrative Science Quarterly*, 52(3), 443-475.

Henderson, R. M., Clark, K. B., (1990). Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. *Administrative Science Quarterly*, 35(1), 9–30.

Kuhn, J. M., Thompson, N. C., (2019), How to measure and draw causal inferences with patent scope. *International Journal of the Economics of Business*, 26(1), 5-38.

Marco, A. C., Sarnoff, J. D. & deGrazia, C. A. (2016). Patent claims and patent scope, USPTO Economic Working Paper No. 2016-04.

McNamee, R. C., (2013), Can't see the forest for the leaves: Similarity and distance measures for hierarchical taxonomies with a patent classification example, *Research Policy*, 42(4), 855-873.

Nelson, R., Winter, S.,(1982), *An evolutionary theory of economic change*. MA: Harvard University Press.

Thompson P & Fox-Kean M. (2005). Patent citations and the geography of knowledge spillovers: a reassessment. *American Economic Review*: 450-460.

Trajtenberg, M. (1990). A penny for your quotes: Patent citations and the value of innovations. The *RAND Journal of Economics*, 21(1), 172-187.

Trajtenberg, M., Henderson, R. & Jaffe, A. (1997). University versus corporate patents: A window on the Basicness of invention, *Economics of Innovation and New Technology*, 5:1, 19-50.

Uzzi, B., Mukherjee, S., Stringer, M. & Jones, B., (2013). Atypical combinations and scientific impact. *Science*, 342(6157), 468-472.

Detecting Multi Word Terms in Patents the same way as Named Entities

Tobias Fink[a], Linda Andersson[b] and Allan Hanbury[c]

[a]*TU Wien, tobias.fink@tuwien.ac.at*
[b]*TU Wien and Artificial Researcher IT GmbH, linda.andersson@tuwien.ac.at* [c]*TU Wien, allan.hanbury@tuwien.ac.at*

ABSTRACT

In English patent document information retrieval, Multi Word Terms (MWTs) are an important factor in determining how relevant a patent document is for a particular search query. Detecting the correct boundaries for these MWTs is no trivial task and often complicated by the special writing style of the patent domain. In this paper we describe a method for detecting MWTs in patent sentences based on a method for detecting technical named entities using deep learning. On our annotated dataset of 22 patents, our method achieved an average precision of 0.75, an average recall of 0.74 and an average F1 score of 0.74. Further, we argue for the use of domain specific word embedding resources and suggest that our model mostly learns whether individual words should be included in MWTs or not.

Keywords: deep learning, multi word term, patent IR, named entity recognition

1. Introduction

Domain specific terminology and technical language often play a key role when determining whether a particular patent document is relevant for a particular search query in Patent Information Retrieval (IR). In English, technical terms of this domain specific terminology are often composed of multiple words making them Multi Word Terms (MWTs), such as “blood cell count”. The meaning of a MWT can be different from the combined meaning of the individual words, which makes it important to detect MWTs as units. When identifying MWTs important words that contribute to the technical nature of the term need to be included and non-technical words need to be excluded. Whether an individual word is an important part of the MWT is not always obvious to the non-expert and might depend on the context of the patent. For example, a “shiny appearance” can be a necessary piece of information in the context of baking products but might be a subjective addition by the author in any other context (see [4]). New MWTs are frequently introduced in the patent domain, be it because of new technology / new concepts that need new MWTs to describe them or be it because of paraphrasing of existing concepts so that the used MWTs refer to a concept more abstractly to widen the scope of a patent claim [6]. As a result, some MWTs that define key-concepts of a technology do not occur very frequently in a patent corpus. In this paper we present a method for detecting MWTs in patent sentences inspired by deep learning methods for detecting keyphrase named entities in scientific text (See [1]). We compare the performance of various model components using a dataset of 22 patents with annotated MWTs. Further, we provide a qualitative analysis of the model performance by looking at the non-training data prediction errors.

2. Multi Word Term Extraction

Since technical terms are often Noun Phrases (NP) ([5]), many methods (such as [3]) require Part-of-Speech (PoS) tagging to detect MWTs. However, [2] note that due to the unique writing style in the patent domain the quality of PoS tagging patent text is problematic, which is why we opt to use a method that does not require PoS tagging to work.
We conduct our experiments on a small dataset of 22 patent documents randomly selected from the CLEF-IP 2013 Topic patent document set. For this dataset we manually annotate the MWT boundaries (i.e. the MWT start and end indices) as they appear in the plain text patent document. Sentences are split into word-token sequences and each word is also split into a sequence of 32 characters. In total, our dataset consists of 232,065 word tokens, 10,337 sentences, and 19,465 MWT instances from a dictionary of 5,099 MWTs. The average MWT dictionary size per patent is 241, while the standard deviation of the MWT dictionary size is 335.
Following the method described in [1], we create a MWT-model architecture (Figure 1) that is designed to transform an input sentence represented as a sequence of words into a BILOU encoded output sequence of labels representing the MWTs in the sentence. The architecture consists of the following components:

- **Word Embedding Component:** consists of a pre-trained word vector which is concatenated to a character representation produced by a small Character-CNN component. We compare domain specific word embeddings with general purpose word embeddings as well as the impact of character representations.
- **LSTM Component:** consists of two Bi-directional LSTM layers.
- **Scoring Component:** produces a sequence of label score vectors, containing a score for each BILOU label.
- **CRF Component:** takes the sequence of label score vectors and predicts the most likely label sequence.

The predicted label sequence is converted to a prediction of MWT boundaries, which are then compared to the

Published in "Proceedings of The 1st Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2019)", Eds. L. Andersson, H. Aras, F. Piroi, A. Hanbury, PatentSemTech 2019, 12 September 2019, Karlsruhe, Germany. https://doi.org/10.34726/pst2019.3| © Authors 2019. CC BY 4.0 License.

An extended version has been submitted to the World Patent Information Journal - Virtual Special Issue: Text Mining and Semantic Technologies in the Intellectual Property domain.

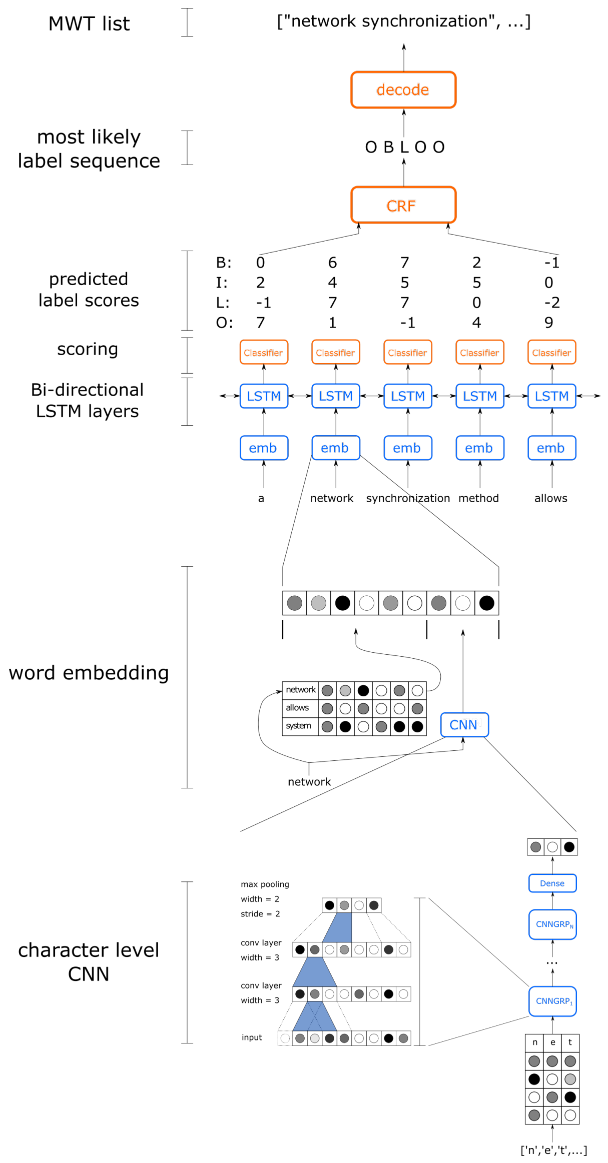


Figure 1: The complete architecture of the MWT model.

ground truth MWT boundaries. To prevent model overfitting we employ early stopping: we keep 10% of our training set patents as validation set and stop training if the validation set F1 score does not improve for a set number of epochs. To measure the performance when detecting MWTs in patent texts, we calculate the precision, recall and F1 score of the model predictions. A MWT prediction counts as a True Positive only if the start and end boundaries exactly match the ground truth boundaries. Further, we provide a qualitative analysis of the model's performance, in particular with respect to prediction errors and their possible causes.

3. Results

Our experiments show that using word embeddings pre-trained on the patent domain outperforms the use of word embeddings pre-trained on Wikipedia and results in an average precision of 0.75, an average recall of 0.74 and an average F1 score of 0.74. In fact, it is necessary to use domain specific word embeddings paired with a character representation produced by the Character-CNN component to perform better than a simple Noun Phrase filter that just annotates all Noun Phrases as MWTs.

Further, we investigate the errors that are made during prediction to get a better idea how the model could be improved. Going through the sentences and the predictions of our best model revealed that the model misses some MWTs by leaving out some words that should be attributed a technical nature, such as “distributed”. This out-of-vocabulary problem might be the result of a too small training set. Sometimes, the model also adds words to MWTs that should not be included, such as non-technical words containing the sub-strings ’activ’ and ’ing’. However, these sub-strings also frequently appear in words that are part of true MWTs, which explains the model’s behaviour.

4. Conclusion

Our experiments suggest that a small dataset of only 22 patents results in an out-of-vocabulary problem and that both a patent specific word embedding resource as well as character representations of words are needed to perform better than basic NP-Filtering. The network appears to learn whether or not individual words or character sequences should be attributed a technical nature, adding them to MWTs if they appear in a MWT context during training or leaving them out if they do not. The same word being included in one MWT but excluded from other MWTs was almost never observed.

By increasing the dataset size it might be possible to reduce the out-of-vocabulary problem in future work. Furthermore, adding additional components, such as a gazetteer or pre-trained language model component, might also improve model performance.

References

[1] Ammar, W., Peters, M., Power, R., Bhagavatula, C., 2017. The AI2 system at SemEval-2017 Task 10 (ScienceIE): semi-supervised end-to-end entity and relation extraction. nucleus 2, e2. URL: https://pdfs.semanticscholar.org/2264/e14e35dc5a3db93437bc408a03171af8c59d.pdf.

[2] Andersson, L., Lupu, M., Palotti, J., Hanbury, A., Rauber, A., 2016. When is the Time Ripe for Natural Language Processing for Patent Passage Retrieval?, in: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, ACM. pp. 1453–1462. URL: http://dl.acm.org/citation.cfm?id=2983858.

[3] Anick, P.G., Verhagen, M., Pustejovsky, J., 2014. Identification of Technology Terms in Patents., in: LREC, pp. 2008–2014. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/701_Paper.pdf.

[4] van Dulken, S., 2014. Do you know English? The challenge of the English language for patent searchers. World Patent Information 39, 35–40.

[5] Justeson, J.S., Katz, S.M., 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. Natural language engineering 1, 9–27.

[6] Nanba, H., Kamaya, H., Takezawa, T., Okumura, M., Shinmori, A., Tanigawa, H., 2009. Automatic translation of scholarly terms into patent terms, in: Proceedings of the 2nd international workshop on Patent information retrieval, ACM. pp. 21–24.

Binary Patent Classification Methods for Few Annotated Samples

Benjamin Meindl[a], Ingrid Ott[b,e], Ulrich Zierahn[c,d]

[a] *University of Lisbon*
[b] *Karlsruhe Institute of Technology (KIT)*
[c] *ZEW Mannheim*
[d] *CESifo Network*
[e] *IfW Kiel*

Abstract

In this paper, we develop binary patent classification algorithms for ambiguous concepts and small sample sizes. These are particularly useful for economic questions, which often require binary classification for implementing ambiguous and subjective concepts, where human classification is time-consuming, so that sample sizes are small. This covers examples such as whether workers are susceptible to automation or not, or whether a device is an automat or not. We compare the performance of naive Bayes, support vector machine, random forest and k-nearest neighbor classifiers with a the spaCy convolutional neural network (CNN) model, as well as spaCy CNN model pre-trained with patent data. The results show overall highest accuracy for the CNN models, with a significantly improved performance through pre-training. Our analysis suggests that the spaCy pre-trained CNN model provides a highly accurate NLP model, feasible for implementation without extensive computation capacity required. Pre-training was particularly beneficial for small sample sizes. Already 100 labeled patents lead to an accuracy of 77.2%. The low sample size required, may encourage researchers in various fields to use manually labeled patent data, for evaluating their specific question.

Keywords: patent classification, small sample size, convolutional neural network, language model pre-training, fast pre-training

1. Introduction

New technologies play a key role for economic development and wealth [1]. This covers a large and currently very active debate on the effects of automation technologies on the labor market [2, 3]. The economic debate often relies on binary classifications to analyze the effects of new technologies on the economy. For example, economists study whether technological change refers to automation or not (e.g. [4]), whether workers are susceptible or non-susceptible to automation (e.g. [5, 6]), how innovation vs. imitation affects the economy (e.g. [7]), or the role of process vs. product innovations for firms (e.g. [8]). Patent texts are well recognized indicators to describe the technological state of the art. As such, patents contain relevant information to measure the mentioned concepts, e.g., by classifying patents that refer to automats vs. non-automats [4]. This is often complex due to the ambiguity of the concepts and the similarity of patents that refer to distinct categories. Being able to assign patents to unique categories allows linking them to other economic data. Until now there only exist few and very broad concordances that allow assigning patents either to technologies [9] or to industries [10]. But these classifications are rather broad.

In this paper, we compare binary patent classifiers, which may be used for analyzing technological change. The main challenge not only lies in the complexity and ambiguity of the concepts, but also in the sample size. Sample sizes are often small, because human coders often require significant time for classifying such cases. These algorithms may be applied to other cases with complex and ambiguous binary classes and few training data.

The rest of this paper is organized as follows: Section 2 provides a description of the underlying patent data and Section 3 our machine learning algorithms. We present and discuss our results in Sections 4. Section 5 concludes.

Published in "Proceedings of The 1st Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2019)", Eds. L. Andersson, H. Aras, F. Piroi, A. Hanbury, PatentSemTech 2019, 12 September 2019, Karlsruhe, Germany. https://doi.org/10.34726/pst2019.4 | *© Authors 2019. CC BY 4.0 License.*

2. Patent Data

We aim at developing a classifier which is able to handle cases with high ambiguity / large overlap. Additionally, it should provide sufficient precision even with low numbers of examples, as hand-classification is costly when human coders have to read large parts of a patent to classify it. In order to develop algorithms which are suited for such cases, we focus on data which contains a binary outcome variable with ambiguous classes. In particular, we rely on patent data, which is particularly suited to study technological change. Moreover, we focus on two selected cooperative patent classification (CPC) classes as our outcome variable to analyze a binary outcome. We focus on two CPC classes which are potentially hard to differentiate for an algorithm in order to train algorithms which are suited for ambiguous cases.

We motivate the choice of our patent sample by the recent interest in robot technologies and the widespread interest this technology field receives in current public and economic debate (e.g., [11, 12, 13]). The United States patent classification (USPC) class 901 - robot - has been mapped to the CPC with the most recent update being from 2012[1]. Most statistically relevant CPC classes related to the USPC class 901 are G 05D, A 61B, G 05B, B 25J, B 23K, B 06B, and G 01N.

Most similar from a technological perspective are CPC classes G 05B and G 05D.[2] We thus restrict our sample to the two sub-classes G 05D and G 05B and use these two classes as a natural delineation to train binary classifiers. G 05D refers to systems for controlling or regulating non-electric variables, e.g., for welding, pressure control, and so on. G 05B relates to control and regulating systems which are "clearly more generally applicable". The fact that G 05B refers to systems which are more generally applicable, whereas G 05D refers to those that control or regulate only non-electric variables, creates a certain ambiguity. Such an abiguity is often present in the economic examples noted above: Without a sufficient training it is often hard to assess for a human, whether a patent is sufficiently generally applicable to be classified as G 05B instead of G 05D. This challenge is similar to the economic samples described in the introduction, such as [4] who define an automat as a device that carries out a process *independently*. Their classification task (i.e., automats vs. non-automats) involves ambiguity, as devices typically require at least some kind of human involvement, so that the interpretation of *independence* remains a subjective assessment of the human coders.

Another objective of the algorithm is to achieve high accuracy with low sample data, as hand-classification is costly when human coders have to read large parts of a patent to classify a patent. [4], for example, build their analysis of patents describing "automats" on 560 hand classified patents. We will compare our algorithms for different sample sizes, to evaluate requirements on sample sizes for potential annotation tasks. We start with the smallest sample size of 100 patents only, which may be mainly relevant for early validation of the feasibility of an idea, and as an input for active learning, which is an early training of the model to select further patents for more efficient classification. Next, we include datasets with 250 and 500 patents. We expect 500 patents to be a potential minimum sample size for analysis, e.g., similar to [4]. Finally we build larger datasets of 1,500 and 5,000 patents, to evaluate the benefit of higher investment of resources for annotation.

We draw our sample data from the USPTO-2m patent abstract dataset [14], which is commonly used for patent classification benchmarking. For each dataset, we draw 50% each G 05D and G 05B examples, whereas patents with both labels are considered as G 05D. For evaluation, we use 250 randomly drawn patents of each category.

3. Patent Classification Algorithms

In our analysis, we compare different approaches for patent classification. [4] use a multinominal naive Bayes (MNB) algorithm to identify patents describing an "automat." Based on 560 manual annotations, they achieve a correct prediction of 80% of patents. One valuable feature of MNB is the ability to interpret results. [4], for example, extract tokens typical for "automats." Support vector machines (SVM) may outperform Naive Bayes [15, 16] or other approaches such as k-nearest neighbor [17] for text classification, and also allow for feature extraction. [18] performed best at the ALTA 2018 patent classification task, using a method based on SVMs.

Further approaches for patent classification are based on neural network (NN) models [19]. [20, 14] describe the potentially high precision of NNs for patent classification and [21] find that they may outperform SVM, particularly for shorter texts. Some recent advances in the field of natural language processing rely on pre-training and fine-tuning NN models (e.g., BERT [22], ULMFiT [23]). [24] outperformed previous approaches

[1] USPC has been deprecated in favor of CPC.

[2] compare https://www.uspto.gov/web/patents/classification/cpc/pdf/us901tocpc.pdf.

of patent classification using patent data to pre-train a BERT convolutional neural network (CNN) model.

Pre-training models such as BERT require extensive computational resources. Therefore, [25, 26] describe alternative models, achieving a significant reduction in computational resource requirements with nearly similar performance. A similar model, called Language Modelling with Approximate Outputs (LMAO) is implemented in the spaCy library[3].

For our analysis, we want to a compare binary classification performance of a pre-trained CNN with alternative approaches. Naive Bayes has been used as a baseline for similar efforts [27]. We use a Bernoulli naive Bayes (BernoulliNB) classifier as a baseline for our work, which accounts particularly for the binary decision. Further, we evaluate an SVM based model, which has been successfully used for various patent classification tasks. Also, we implement a random forest classifier (RandomForest) and a k-nearest neighbor classifier (k-NN) for comparison.

BernoulliNB, SVM, RandomForest, and K-NN classifiers are implemented using Scikit-learn. Therefore, we lemmatize words (using NLTK[4]), remove stopwords, and extract the most relevant words per document through term frequency-inverse document frequency scores (TF-IDF), using unigrams as well as bigrams. [28] finds that TF-IDF analysis using bigrams (instead of unigrams only) may lead to higher accuracy, as it accounts for complex multi-word expressions. We use the Scikit-learn model selection, GridSearchCV, for optimization of model parameters.

We implement a CNN based classifier using spaCy, which is a library aiming at providing a combination of high accuracy and speed. This is especially relevant for patent classification, as it enables research on large patent data sets with reasonable resources. Further, it allows resource efficient LMAO pre-training for patent specific context.

Our analysis includes two spaCy based approaches. First, we use the default large English language model. Second, we use the same model pre-trained with patent data (we refer to it as spaCy$_{\text{pre}}$). To assure high contextual relevance of pre-training, we use the 25,212 patents in the class G 05 from the USPTO-2m dataset. The algorithm ran 200 passes over the dataset until the loss function did not further decrease. In addition, we run the same models with the software prodigy[5]. Prodigy builds on spaCy and allows for straightforward implementation of natural language processing analysis and annotation. It provides a simple API requiring only basic knowledge in programming. We want to evaluate whether using the tools compromises performance compared to a manual implementation of spaCy.

4. Results and Discussion

A comparison of the different algorithms shows that the pre-trained CNN model outperforms remaining models (see table 1) for each sample size. The regular spaCy model performs second best for all sample sizes. From the remaining models, the BernoulliNB classifier performed best for all sample sizes but the largest one. The performance of the SVC model fluctuated strongly for different sample sizes, and did even decrease, e.g., comparing the 1,500 dataset with the 250 dataset. RandomForest and k-NN were within lowest performing classifiers for all sample sizes, however, they reach a reasonable accuracy for the largest dataset. We thus find that the pre-trained CNN model performs best as a binary patent classifier for hard-to-classify concepts.

The results further show a significant increase in performance through pre-training with patent data. The benefits are strongest for small sample sizes, where 100 annotations led to accuracy scores of 77.2%, compared to a score of 72.5% for the CNN without pre-training. This score suggests, that pre-trained neural network may be well suitable for active learning, which aims at increasing the efficiency of annotations through active learning [29]. The performance advantage of pre-training, however, decreases with sample size and almost disappears for the largest data set. Accordingly, we find that pre-training is particularly useful for small data sets, but provides negligible performance advantages with large data sets of around 5,000 or more annotated samples. Future research may evaluate, whether more expensive pre-training methods provide even stronger models.

Our best-performing CNN achieves an accuracy of 0.832 and 0.866 with sample sizes of 500 and 1500 patents. These accuracy scores may be appropriate for a number of further analyses and may encourage future researchers to use labeled patent data for their analyses.

Moreover, the spaCy LMAO pre-training does not require extensive computation capacity. Therefore, the described methods may be suitable for a broad range of researchers, providing high accuracy and enabling efficient implementation.

In addition to the results shown in the table, we ran the spaCy models through the Prodigy software. The

[3] https://spacy.io/

[4] https://www.nltk.org

[5] https://prodi.gy

Model	**Sample size**				
	100	250	500	1,500	5,000
BernoulliNB	0.706	0.776	0.798	0.808	0.842
SVC	0.612	0.536	0.794	0.774	0.858
RandomForest	0.590	0.668	0.752	0.770	0.836
K-NN	0.598	0.704	0.716	0.772	0.838
spaCy	0.726	0.786	0.806	0.858	0.872
spaCy$_{\mathbf{pre}}$	**0.772**	**0.800**	**0.832**	**0.866**	**0.874**

Table 1: Comparison of patent classification performance. The models implemented are Bernoulli naive Bayes (BernoulliNB), support vector machine (SVC), random forest, k-nearest neighbour, spaCy large English model, and a spaCy model pre-trained with patent data. The models have been tested with different sample sizes, of 100, 250, 500, 1,500, and 5,000 patents in categories G 05D, and G 05B. Scores relate to recognition of G 05D.

results were similar to both spaCy models and are thus not listed in Table 1. This implies that relying on a simple API that requires only basic knowledge in programming comes at little performance costs, rendering the methods proposed in this paper potentially accessible to researchers from disciplines with typically less training in programming, such as e.g. economists.

5. Conclusions

Patent classification, in general, is an active research field. Besides pre-classification of patent applications, which is highly relevant for patent offices [17], also other fields may benefit from advances in this area. Particularly economists may benefit from improved methods of patent analyses. [30], for example, describe the lack of high-quality data and empirically informed models as a key challenge for a better understanding of automation technologies. Patent data may be a rich source of data to address this challenge.

Our work contributes to patent as well as NLP research by evaluating a powerful pre-trained CNN based approach for binary patent classification. The proposed method offers a fast, high accuracy tool enabling a broad range of researchers conducting patent classification or other text classification tasks. We find that pre-training significantly raises performance particularly in small samples of annotated data, while the performance surplus declines for larger samples.

We further find that the methods provide a high accuracy, do not require high computational resources, and that relying on Prodigy as a simple API does not result in noticeable performance losses. This implies that the methods proposed here are both useful and potentially accessible to researchers from other disciplines.

Competing Interests

We declare that we have no significant competing financial, professional, or personal interests that might have influenced the performance or presentation of the work described in this manuscript.

References

[1] D. Acemoglu, Introduction to Modern Economic Growth, Princeton University Press, 2009.

[2] J. Mokyr, C. Vickers, N. L. Ziebarth, The history of technological anxiety and the future of economic growth: Is this time different?, Journal of Economic Perspectives 29 (3) (2015) 31–50.

[3] D. Autor, Why are there still so many jobs? the history and future of workplace automation, Journal of Economic Perspectives 29 (3) (2015) 3–30.

[4] K. Mann, L. Püttmann, Benign effects of automation: New evidence from patent texts (2018).

[5] C. B. Frey, M. A. Osborne, The future of employment: How susceptible are jobs to computerization?, Technological Forecasting and Social Change 114 (2017) 254–280.

[6] M. Arntz, T. Gregory, U. Zierahn, Revisiting the risk of automation, Economics Letters 159 (2017) 157–160.

[7] P. S. Segerstrom, Innovation, imitation, and economic growth, Journal of Political Economy 99 (4) (1991) 807–827.

[8] A. Bartel, C. Ichniowski, K. Shaw, How does information technology affect productivity? Plant-level comparisons of product innovation, process improvement, and worker skills, Quarterly Journal of Economics 122 (4) (2007) 1721–1758.

[9] U. Schmoch, Concept of technology classification for country comparisons, Final report to the World Intellectual Property Organisation (WIPO), Fraunhofer Institute for Systems and Innovation Research (ISI) (2008).

[10] eurostat, Patent statistics: Concordance ipc v8 - nace rev.2, Tech. rep., Eurostat (2014).
URL `https://circabc.europa.eu/sd/a/d1475596-1568-408a-9191-426629047e31/2014-10-16-Final%20IPC_NACE2_2014.pdf`

[11] D. Acemoglu, P. Restrepo, Robots and jobs: Evidence from US labor markets, NBER Working Paper 23285 (2017).

[12] W. Dauth, S. Findeisen, J. Südekum, N. Wössner, German robots - the impact of industrial robots on workers, IAB Working Paper 30/2017.

[13] G. Graetz, G. Michaels, Robots at work, Review of Economics and Statistics 100 (5) (2018) 753–768.

[14] S. Li, J. Hu, Y. Cui, J. Hu, DeepPatent: patent classification with convolutional neural networks and word embedding, Scientometrics 117 (2) (2018) 721–744. doi:10.1007/s11192-018-2905-5.

[15] T. Joachims, Text Categorization with Support Vector Machines : Learning with Many Relevant Features, European conference on machine learning (1998) 137–142.

[16] C. J. Fall, A. Törcsvári, K. Benzineb, G. Karetka, Automated Categorization in the International Patent Classification, Acm Sigir Forum 37 (1) (2003) 10–25.

[17] M. Krier, F. Zacc, Automatic categorisation applications at the European patent office, World Patent Information 24 (2002) 187–196.

[18] F. Benites, S. Malmasi, M. Zampieri, Classifying Patent Applications with Ensemble Methods, Proceedings ofAustralasian Language Technology Association Workshop (2018) 89–92.

[19] A. Abbas, L. Zhang, S. U. Khan, A literature review on the state-of-the-art in patent analysis, World Patent Information 37 (2014) 3–13. doi:10.1016/j.wpi.2013.12.006.

[20] M. F. Grawe, C. A. Martins, A. G. Bonfante, Automated Patent Classification Using Word Embedding, 16th IEEE International Conference on Machine Learning and Applications (2017) 408–411doi:10.1109/ICMLA.2017.0-127.

[21] W. Zaghloul, S. M. Lee, S. Trimi, Text classification: neural networks vs support vector machines, Industrial Management & Data Systems 109 (5) (2009) 708–717. doi:10.1108/02635570910957669.

[22] J. Devlin, M. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 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) (2019) 4171–4186.
URL https://www.aclweb.org/anthology/N19-1423

[23] J. Howard, S. Ruder, Universal Language Model Fine-tuning for Text Classification, Proceedings ofthe 56th Annual Meeting ofthe Association for Computational Linguistics (Volume 1: Long Papers) (2018) 328–339.

[24] J. Lee, J. Hsiang, PatentBERT: Patent classification with fine-tuning a pre-trained BERT model (2019). arXiv:1906.02124.
URL http://arxiv.org/abs/1906.02124v2

[25] L. H. Li, P. H. Chen, C.-J. Hsieh, K.-W. Chang, Efficient contextual representation learning without softmax layer (2019). arXiv:1902.11269.

[26] S. Kumar, Y. Tsvetkov, Von mises-fisher loss for training sequence to sequence models with continuous outputs (2018). arXiv:1812.04616v3.

[27] D. Mollá, D. Seneviratne, Overview of the 2018 ALTA shared task: Classifying patent applications, in: Proceedings of the Australasian Language Technology Association Workshop 2018, Dunedin, New Zealand, 2018, pp. 84–88.
URL https://www.aclweb.org/anthology/U18-1011

[28] E. D'hondt, S. Verberne, C. Koster, L. Boves, Text Representations for Patent Classification, Computational Linguistics 39 (3) (2013) 755–775. doi:10.1162/COLI.

[29] S. Tong, D. Koller, Support Vector Machine Active Learning with Applications to Text Classification, Journal of Machine Learning Research (2001) 45–66.

[30] M. R. Frank, D. Autor, J. E. Bessen, E. Brynjolfsson, M. Cebrian, D. J. Deming, M. Feldman, M. Groh, J. Lobo, E. Moroa, D. Wang, H. Youn, I. Rahwan, Toward understanding the impact of artificial intelligence on labor, PNAS 116 (14) (2019) 6531–6539. doi:10.1073/pnas.1900949116.

Deep Learning based Pipeline with Multichannel Inputs for Patent Classification

Mustafa Sofean

FIZ Karlsruhe, Hermann-von-Helmholtz-Platz 1. 76344 Eggenstein-Leopoldshafen
Mustafa.Sofean@fiz-karlsruhe.de

Abstract

Patent document classification as groundwork has been a challenging task with no satisfactory performance for decades. In this work, we introduce a deep learning pipeline for automatic patent classification with multichannel inputs based on LSTM and word vector embeddings. Sophisticated text mining methods are used to extract the most important segments from patent texts, and a domain-specific pre-trained word embeddings model for the patent domain is developed; it was trained on a very large dataset of more than five million patents. A deep neural network model is trained with multichannel inputs namely embeddings of different segments of patent texts, and sparse linear input of different metadata. A series of patent classification experiments are conducted on different patent datasets, and the experimental results indicate that using the segments of patent texts as well as the metadata as multichannel inputs for a deep neural network model, achieves better performance than one input channel.

Keywords: Patent Analysis, Neural Network, Deep Learning, Patent Classification

1. Methods

Patent classification is a kind of knowledge management where documents are assigned into predefined categories. Due to the extremely complicated patent language and hierarchical patent classification scheme, many previous studies focused only on whole texts of patent or some general sections such as title, abstract, detailed description and claims [2] [1]. They did not consider the most important sections like background, technical field, summary, and independent claims that need specific text mining tools to extract.

1.1. Semantic Structure of patent and Embeddings

Efficient text mining services are used for semantic structuring of the patent texts [3]. The first service is used to structure the description part of patent text into structured segments such as the technical field, background, summary, and the embodiments [5]. The second service is able to automatically identify the complete claim hierarchy within patent texts [4]. In addition, a domain-specific word and phrase embeddings model is developed for the patent domain. The model is trained on more than five million patent documents and can be used for word/phrase similarity or patent analysis such as classification tasks.

1.2. Deep Learning based Pipeline Architecture

Firstly, we extract the most important segments of patent texts which are title, abstract, technical field, background, summary, and the independent claim. For texts of each segment, a tokenization process is used for breaking the text into individual words, and the sequence length of each segment is set according to the maximum length of each. The deep learning architecture has two components: deep, and wide. It feed-forward neural networks with embeddings of each segment, and uses them as deep layers for deep neural network model, and the patent metadata on the other hand is used as a wide part for the model. Specifically, the architecture is described as follows: for the wide components of the model, we used one-hot representation for patent metadata features (such as inventors, citations, and assignees), these one-hot vectors are fed into separate sub-networks, and

*Published in "Proceedings of The 1**st Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2019)**", Eds. L. Andersson, H. Aras, F. Piroi, A. Hanbury, PatentSemTech 2019, 12 September 2019, Karlsruhe, Germany. https://doi.org/10.34726/pst2019.5 | © Authors 2019. CC BY 4.0 License.*

An extended version has been submitted to the World Patent Information Journal - Virtual Special Issue: Text Mining and Semantic Technologies in the Intellectual Property domain.

Table 1: Evaluation Results. (TI:title, AB:abstract, TECHF: Technical Fields, BACK: Background, SUMM: Summary, IND_CLAM: Independent Claim, INVs: Inventors, and PAs: Patent Assignees)

Input	**Accuracy**	**Precision**	**Recall**	**F1-score**
All texts of segments as one channel	67%	84%	61%	71%
TI, AB, TECHF , BACK, SUMM, and IND_CLAM as multichannels	**74%**	**92%**	**63%**	**75%**
TI and TECHF as multichannel inputs	66%	83%	59%	69%
TI and TECHF, INVs, and PAs as multichannel inputs	68%	85%	61%	71%

at the end they are represented as deep networks. For the deep components of the model, deep layers are created for the most important patent text segments. These are sequential input to a Long Short-Term Memory (LSTM) network that takes the embeddings as inputs that are obtained by using a pre-trained word embeddings model to encode each segment texts into vectors, and then we feed them into LSTM layers. To avoid network overfitting and help network stability, additional layers are added for each input channel, dropout layer is used to drop out 30% of input in order to prevent neural networks from overfitting, and Batch normalization layer is used to normalize the input layer by adjusting and scaling the activations. The exponential linear unit (ELU) is used as activation function. Finally, we concatenated nine components which are text-based LSTM layers (title LSTM, abstract LSTM, technical field LSTM, background LSTM, summary LSTM, and independent claim LSTM), and metadata-based LSTM layers (inventors, assignees, and citations) into a final set of deep layers with dropout, batch normalization, and softmax activation function for multi-class and sigmoid for multi-label classification task.

1.3. Experimental Results

The dataset in this work is extracted from databases of the European Patent Office (EPO) and the World Intellectual Property Organization (WIPO). All extracted patents contain the title, abstract, detailed description, claims, and at least one IPC label. The total number of extracted records in the dataset is about 1,915,308 patents filed between 1978 and 2016. The segmentation tools [3] [4] were used to extract the most important sections (technical field, background, summary of invention and independent claim from patent texts. All patent documents are classified into related subclass level of IPC, and we used four evaluation measures namely accuracy, precision, recall, and F1. A series of patent classification experiments are conducted on the dataset, and we also studied how the full text, different parts of a patent information, and their combination affect the classification performance. The evaluation results are shown in the table 1. The best performance we obtained is 74%, 92%, 63%, and 75% for accuracy, precision, recall, and F1, respectively. The result in this work indicates that using the segments of patent text as multichannel inputs improved the performance of patent classification in terms of all evaluation criteria.

2. Conclusion

In this work, we introduced a deep learning based pipeline for large-scale patent classification. Different parts of patent information are used as multichannel inputs for a Long Short-Term Memory (LSTM) that takes the both vectors (embeddings and one-hot) in order to learn a patent classification model. The experimental results indicated that using the segments of patent texts as well as the metadata as multichannel inputs for a deep neural network model, achieve a good performance.

3. References

[1] Assad Abbas, Limin Zhang, S.U.K., 2015. A literature review on the state-of-the-art in patent analysis. World Patent Information 37, 3–13.

[2] Juan Carlos Gomez, M.F.M., 2014. A survey of automated hierarchical classification of patents. Springer International Publishing .

[3] Mustafa Sofean, Hidir Aras, A.A., 2018. A workflow-based large-scale patent mining and analytics framework. Proceedings of 24th International Conference on Information and Software Technologies (ICIST); .

[4] Rene Hackl-Sommer, M.S., 2015. Patent claim structure recognition. Archives of Data Science .

[5] Sofean, M., 2017. Automatic segmentation of big data of patent texts. International Conference on Big Data Analytics and Knowledge Discovery. DaWaK , 343–351.

Semantic Views – Interactive Hierarchical Exploration for Patent Landscaping

Tatyana Skripnikova[1]

KIT - Karlsruhe Institute of Technology

Hidir Aras[2]

FIZ Karlsruhe - Leibniz Institute for Information Infrastructure

Anna Weißhaar, Sebastian Blank, Hans-Peter Zorn[3]

inovex GmbH

Abstract

In this paper, we investigate whether a semantic representation of patent documents provides added value for a multi-dimensional visual exploration of a patent landscape compared to traditional approaches. Word embeddings from a pre-trained model created from patent text are used to calculate pairwise similarities for representing each document in the semantic space. Then, a hierarchical clustering method is applied to create several semantic aggregation levels for a collection of patent documents. For visual exploration, we have seamlessly integrated multiple interaction metaphors that combine semantics and additional metadata for improving hierarchical exploration of large document collections.

Keywords: Patent landscaping, word embeddings, hierarchical clustering, text analysis, semantic exploration, multi-dimensional exploration, visual user interface.

1. Introduction

The number of written works describing scientific progress is steadily increasing, which necessitates the development of supportive tools for their efficient analysis. Developing a visualization approach to facilitate the examination proves to be a challenging task. This is due to the complexity of these documents, which are not only characterized by their textual content, but also by a number of metadata attributes of various kinds, including information about relationships between them.

Patent landscaping [1] is an example of a task in which complex document explorations take place. With the help of patent landscaping, companies acquire competitive advantages and steer their research and development efforts. However, with hundreds to thousands of patent documents that have to be considered per patent landscaping report, human perception must be aided in the task of finding patterns in data to prevent cognitive overload.

We propose an approach that allows for a multi-dimensional visual exploration [2] based on both semantics and metadata from the patent documents. Semantic embeddings [3] are widely used in natural language processing to capture relationships between text documents. Nonetheless, when trying to visualize those relationships, we face the problem that positions and distances in the embedding

[1] https://www.kit.edu

[2] https://www.fiz-karlsruhe.de

[3] https://www.inovex.de

Published in "Proceedings of The 1st Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2019)", Eds. L. Andersson, H. Aras, F. Piroi, A. Hanbury, PatentSemTech 2019, 12 September 2019, Karlsruhe, Germany. https://doi.org/10.34726/pst2019.6 — Authors 2019. CC BY 4.0 License.

An extended version has been submitted to the World Patent Information Journal - Virtual Special Issue: Text Mining and Semantic Technologies in the Intellectual Property domain.

space are not easily explainable and can hardly be understood by themselves. As for creating a patent landscape the question what is "in/out" of a focused topical region is crucial, we utilize semantic similarity of documents for creating a patent landscape [4] followed by clustering [5] the documents at 3 aggregation levels employing hierarchical agglomerative clustering.

2. Semantic Exploration of a Patent Landscape

In order to semantically explore patent documents, two challenges exist which must be reflected in our approach:

1. visually presenting high-dimensional *semantic representations* of documents in a way that is intuitively understood, and
2. supporting *semantic interactions*, which means that the display adapts to the intentions of the user with regard to information density and level of detail [6].

The user interface and the interaction metaphors it offers are designed to handle these challenges by utilizing a number of coordinated views which respond to the changes in each other's states. The scatter plot is the main area of the visualization representing the semantic space. At the same time, the histogram and sunburst views display metadata attributes from the dataset in an aggregated form. They enable filtering and highlighting of the data across all views via *brushing and linking*, which means that "the change to the representation in one view affects the representation in the other views as well" [7].

The interactions connecting the views fall into one of three groups: selection, highlighting and resetting the current selection. The implementation is consistent across all views: 1) clicking means selecting an object/group, 2) hovering with the mouse causes a highlighting of an object/group which is a preview of the selection, 3) clicking on the background of a view resets the selection.

3. Evaluation

In order to evaluate the question of how semantic information in combination with rich metadata can be used to enhance the visual exploration of large document collections, we conducted a summative study in form of a think-aloud-experiment with several patent experts. The subjects of the study are employees of FIZ Karlsruhe and have acquired significant experience over the years with patent matters in general and with patent research and patent landscaping in particular. This makes them appropriate candidates to study the complex cognitive processes that happen during the exploration.

We evaluated the visualization approach by means of tasks the users had to perform and by means of questionnaires for capturing user feedback.

The usability study indicates that visualization metaphors and interaction techniques were appropriately chosen.

4. Conclusion

We set out to present a novel approach for the hierarchical exploration of large document collections combining semantics and additional metadata. Our research shows that the chosen interaction techniques are consistent and intuitive. The proposed visualization approach provides added value for patent landscaping, and can also be applied to other document exploration tasks.

References

[1] K. W. Boyack, B. N. Wylie, G. S. Davidson, Domain visualization using vxinsight for science and technology management, Journal of the American Society for Information Science and Technology 53 (9) (2002) 764–774. doi:10.1002/asi.10066.

[2] K. Wittenburg, G. Pekhteryev, Multi-Dimensional Comparative Visualization for Patent Landscaping, in: IEEE VIS Workshop 2015 BusinessVis15.

[3] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representationsarXiv:1802.05365.
URL http://arxiv.org/abs/1802.05365

[4] A. Abood, D. Feltenberger, Automated patent landscaping, Artificial Intelligence and Law 26 (2) (2018) 103–125. doi:10.1007/s10506-018-9222-4.

[5] A. Skupin, A Cartographic Approach to Visualizing Conference Abstracts, IEEE Computer Graphics and Applications 22 (1) (2002) 50–58. doi:10.1109/38.974518.

[6] D. Modjeska, Navigation in Electronic Worlds: Research Review for Depth Oral Exam (May) (1997) 1–56.

[7] Marti A. Hearst, Modern Information Retrieval, Addison-Wesley-Longman Publishing co., 1999. doi:10.7748/ns.4.42.26.s69.

Visual Programming for Patent Mining

Farag Saad, Hidir Aras[1]

FIZ Karlsruhe Leibniz Institute for Information Infrastructure

Stefan Helfrich[2]

KNIME GmbH

Abstract

In this paper we describe how to employ and extend the KNIME Analytics Platform with a Hadoop backend in order to realize scalable text mining workflows for annotating and linking large-scale patent data.

Keywords: Patent Mining, KNIME Analytics Platform, Workflows, Visual Programming, Hadoop, Scalability

1. Introduction

The aim of scientific workflow systems is to allow users to systematically describe scientific processes or methods, e.g. for data analysis. For example, scientists from the life sciences need support how to perform annotation tasks on large amounts of chemical texts. Still today, programming or scripting skills are not necessarily part of a life scientists tool built to accomplish their analysis tasks. Hence, a platform which allows to remedy this by means of visual programming (instead of requiring a specific programming language) is needed.

The KNIME Analytics Platform [1] is an easy to use and comprehensive open source data integration, analysis, and exploration platform, designed to handle large amounts of heterogeneous data. It implements intuitive usage concepts and allows users to perform programming tasks in a visual manner: the data flow and processing steps in a workflow are modelled by inserting and connecting modules (so called “nodes”) as shown in Figure 1. Furthermore, KNIME allows for easy configuration and execution of nodes in its graphical user interface. It provides over 2,000 native nodes that are continually developed and maintained by KNIME AG. These core nodes cover generic functionality that is independent of the underlying data. The most prominent examples are statistical evaluation, data mining and machine learning, as well as (customizable) interactive visualizations. Many more of these nodes are provided through open source integrations, e.g. Apache Spark or Apache Hadoop for big data processing, H2O for high performance machine learning, Python and R for scripting and plotting, and extensions including image processing, cheminformatics, or bioinformatics.

The KNIME Text Processing Extension, in particular, was designed and developed to read and process textual data, and transform it into numerical data (document and term vectors) to, for example, apply data-agnostic nodes from KNIME Analytics Platform. The extension enables reading, processing, mining and visualization of textual data in a convenient way. Processing may involve, among others, recognition and tagging of named entities, filtering (e.g. named entity filters), and stemming. Frequencies of words can be computed, keywords can be extracted, and documents can be visualized (e.g. tag clouds), among other things.

Despite the fact that KNIME provides a diverse set of built-in nodes, there is a high demand to extend the platform’s functionalities in order to fulfill specific requirements. This is taken into account and the platform allows to develop and inte-

[1] https://www.fiz-karlsruhe.de
[2] https://www.knime.com

Published in "Proceedings of The 1st Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2019)", Eds. L. Andersson, H. Aras, F. Piroi, A. Hanbury, PatentSemTech 2019, 12 September 2019, Karlsruhe, Germany. https://doi.org/10.34726/pst2019.7 | © Authors 2019. CC BY 4.0 License.

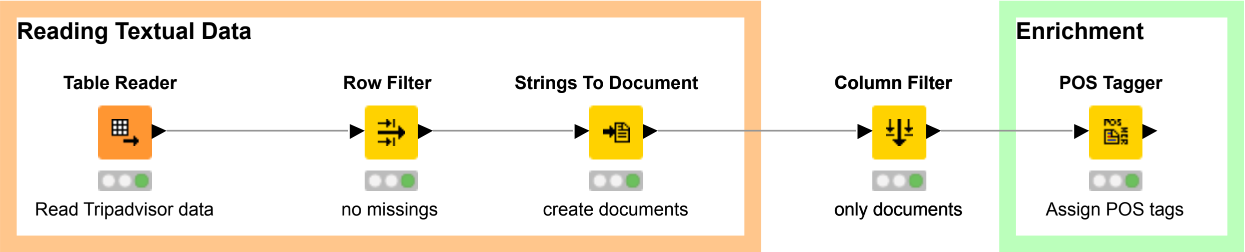


Figure 1: Simple workflow for POS Tagging implemented with KNIME Analytics Platform.

grate custom nodes for implementing more complex scientific workflows, complementing KNIME's core functionality.

In this paper, we take advantage of this extensibility feature by developing new nodes for building scalable patent text mining workflows with the KNIME Analytics Platform. To be able to build such workflows one has to

- implement custom nodes to be used for visual programming in order to execute annotation tasks and interact with the results, and
- integrate a scalable service processing backend, e.g. based on Hadoop/MR2.

A proof of concept implementation for annotating and linking chemical entities shows that our approach allows for creating efficient, scalable, and easy to use patent mining workflows by means of visual programming.

2. A Patent Text Mining Use Case

In the following, we describe a use case from patent text mining [2] for annotating and linking of chemical entities in patent texts. As a prerequisite to realize a dedicated scalable annotation workflow, several custom KNIME nodes for essential tasks have been developed and integrated.

2.1. Scalable Annotation Workflows

In recent years, a strong increase in publications (patents and scientific articles) related to the life sciences has been observed. The extraction of meaningful information e.g. named entities from these publications is no longer achievable within a reasonable time without big data processing. Therefore, a sophisticated distributed system such as Hadoop can be employed in order to perform extraction, transformation and loading tasks (ETL) efficiently. However, the integration of such a scalable system into customized workflows e.g. as KNIME workflows, is often a challenging task. In order to allow for scalable processing with a customized workflow, three processes need to be initiated. First, the subset of the input data has to be selected and necessary information sent to a KNIME node which is responsible for distributed processing. Secondly, a KNIME node which allows for distributed data processing has to be executed. And thirdly, the annotated data then needs to be retrieved, post-processed and visualized for user interaction.

In the following, we describe our approach for performing scalable annotation tasks in technical detail, with a robust annotation and visualization service integrated into the KNIME analytics platform. In order to enable scalable processing and mining in KNIME workflows, four custom nodes have been developed and integrated seamlessly.

A scalable annotation workflow comprises steps for preparing the execution of an annotation task, e.g. setting configuration parameters, creating database tables, deployment to a multi-node cluster, and, finally, its distributed execution. Figure 2 shows the *PhoenixUnzipData* custom node which is used to select the ids of a subset of patent documents for annotation and its configuration e.g. entering the SQL statement using the *Table Creator* node and handing over as parameter.

The obtained id list serve as input for the *Phoenix_OP* custom node, which is shown in Figure 2. Herewith, a database table for storing the id list is created and all ids inserted. Besides that, it is needed to specify the table name for storing the annotations, which is again done via a Table Creator node. At the same time the table name serves as an input parameter for the *GenericServiceRunner* node for the distributed processing in the Hadoop cluster. This custom KNIME node uses

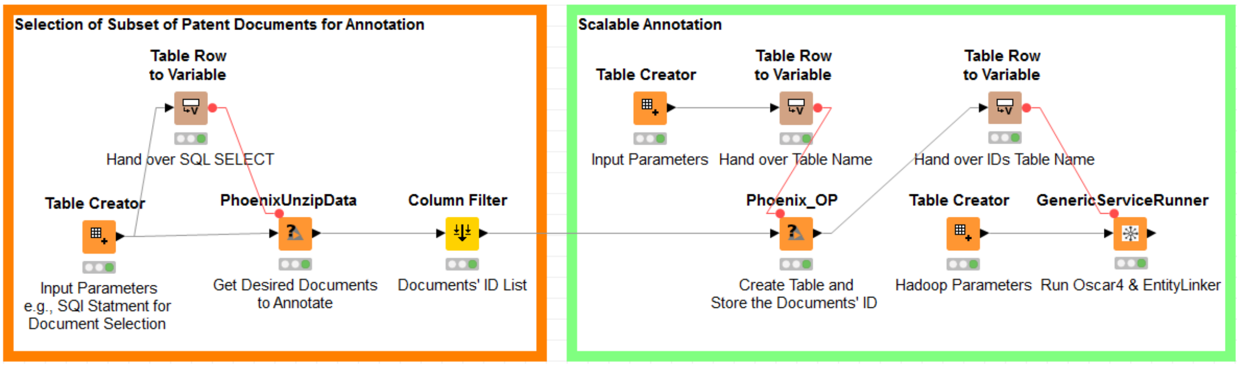


Figure 2: Scalable Chemical Annotation Workflow.

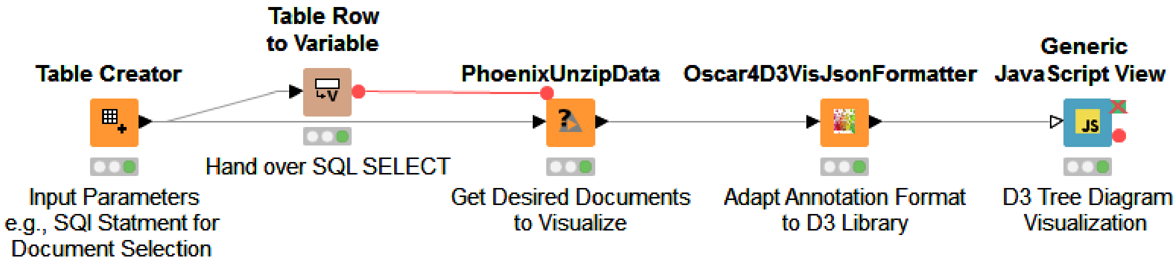


Table Creator
Input Parameters e.g., SQl Statment for Document Selection
Table Row to Variable
Hand over SQL SELECT
PhoenixUnzipData
Get Desired Documents to Visualize
Oscar4D3VisJsonFormatter
Adapt Annotation Format to D3 Library
Generic JavaScript View
D3 Tree Diagram Visualization

Figure 3: Visualization Workflow for Chemical Annotations.

the Hadoop/YARN REST API in order to communicate with the cluster to execute a dedicated Hadoop/MR2 job.

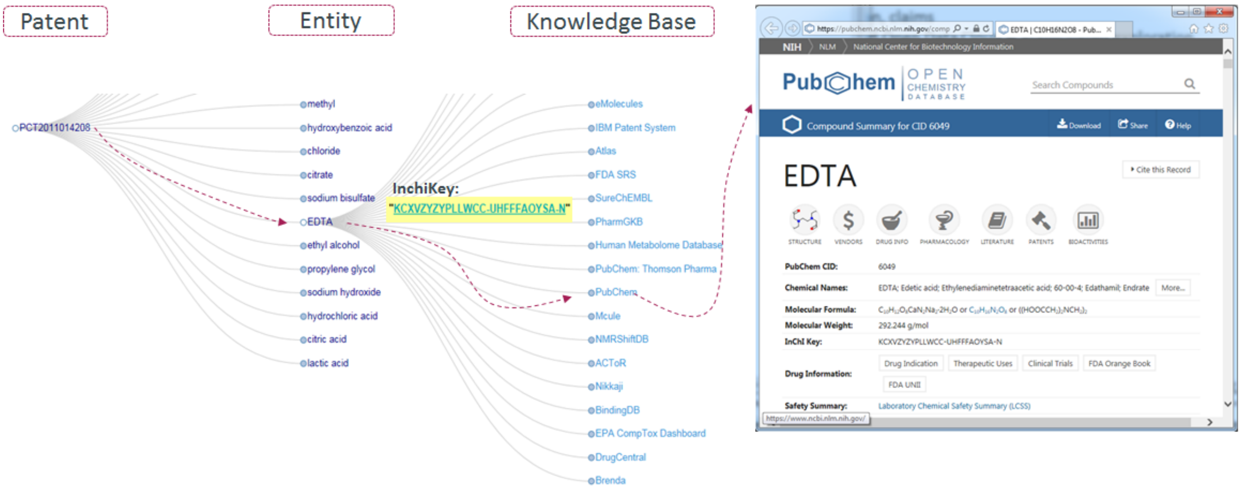
Before a scalable annotation service can be executed for large-scale data, the java library (jar) of the service must be deployed to the Hadoop cluster manually and stored in the distributed file system HDFS. Alternatively, the HDFSUpload node from the KNIME Big Data Extensions could also be used.

In our use case a service for annotating and linking chemical entities employing the OSCAR4 chemical tagger [3] and the *EntityLinker* service (see Section 2.2) was developed and used for the workflow. Once the annotation service and the needed resources are available on the cluster, it can be horizontally scaled and executed via the *GenericServiceExecutor* Hadoop/MR2 job. This is possible, because the service implements a standard java interface, which is seamlessly integrated into the Hadoop service executor job. The results of the annotation service, i.e. chemical entities and links to external resources such as to ChEMBL, DrugBank, etc.for each processed patent in the selected set are stored to the Phoenix database and can be accessed for later processing or visualization.

2.2. Linking Chemical Entities via Chemical Structure Identifiers

The mapping between annotated chemical entities and external resources is performed by the EntityLinker using the *UniChem* [4] web services. UniChem efficiently produces cross-references between chemical structure identifiers (e.g. InCHI Key) from different databases and includes the entities' identifiers of 28 databases such as ChEMBL, DrugBank, IBM Patent System, ChEBI, SureChEMBL, PharmGKB, NIH Clinical Collection, etc. In our use case, the output of the GenericServiceExecutor job (stored in a previously created Phoenix table) includes the annotations along with the external resource links and the status (successful, error) for each patent document.

https://www.ebi.ac.uk/unichem/info/webservices



Compound Summary for CID 6049

EDTA

PubChem CID: 6049

Chemical Names: EDTA; Edetic acid; Ethylenediaminetetraacetic acid; 60-00-4; Edathamil; Endrate

Molecular Formula: $C_{10}H_{12}O_8CaN_2Na_2 \cdot 2H_2O$ or $C_{10}H_{16}N_2O_8$ or $((HOOCCH_2)_2NCH_2)_2$

Molecular Weight: 292.244 g/mol

InChI Key: KCXVZYZYPLLWCC-UHFFFAOYSA-N

Drug Information: Drug Indication | Therapeutic Uses | Clinical Trials | FDA Orange Book | FDA UNII

Safety Summary: Laboratory Chemical Safety Summary (LCSS)

Figure 4: D3 Tree Diagram Visualizations

2.3. Visualization of Results

In the last part of the scalable annotation workflow shown in Figure 3), the PhoenixUnzipData custom node is used to select and visualize a subset of the annotated documents. The *Oscar4D3VizJsonFormatter* custom node is used to convert the resulting annotations to the format of the D3.js visualization library. In the example shown in Figure 3, the annotations of a patent document are visualized as a "D3 Tree Diagram". KNIME's *Generic JavaScript View* node is used for implementing and running the D3 Tree Diagram code.

The resulting tree diagram visualization gives the user excellent view and shows the annotated document with its chemical entities linked to their corresponding external knowledge resources. Using the tree diagram visualized annotation, the user can easily by mouse click navigate between entities and their external knowledge resources. For example, as Figure 4 shows by clicking on the annotated entity EDTA, the workflow will react by displaying a list of the linked external resources e.g. PubChem. For example, when the user clicks on PubChem, the workflow will react and display detailed information about the annotated entity EDTA from the PubChem database e.g. PubChem id, InChi Key, Drug information, Molecular formula etc.

3. Conclusion

The KNIME Analytics Platform with its extensions for e.g. text processing allows for building patent mining workflows in a visual fashion. The flexibility and openness of the platform makes it a perfect candidate for rapid development and deployment of large-scale analysis and mining workflows without vendor lock in.

We have shown an example that employs custom KNIME nodes and a scalable service infrastructure based on Hadoop in order to enrich and link chemical entities in patent text to external knowledge sources. Furthermore, we have shown how the user can interact with the visualized result by navigating between linked external resources and exploring new information related to the annotated chemical entities such as InChi Key, SMILES, molecular formula, molecular weight etc.

References

[1] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel, KNIME: The Konstanz Information Miner, in: Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007), Springer, 2007.

[2] M. Sofean, H. Aras, A. Alrifai, A workflow-based large-scale patent mining and analytics framework, in: R. Damaševičius, G. Vasiljevienė (Eds.), Information and Software Technologies, Springer International Publishing, Cham, 2018, pp. 210–223.

[3] D. M. Jessop, S. E. Adams, E. L. Willighagen, L. Hawizy, P. Murray-Rust, Oscar4: a flexible architecture for chemical text-mining, Journal of Cheminformatics 3 (1) (2011) 41.

[4] J. Chambers, M. Davies, A. Gaulton, A. Hersey, S. Velankar, R. Petryszak, J. Hastings, L. Bellis, S. McGlinchey, J. Overington, Unichem: A unified chemical structure cross-referencing and identifier tracking system, Journal of cheminformatics 5 (2013) 3.

Deep Learning Services for Patents

Mustafa Sofean and Ahmad Alrifai

FIZ Karlsruhe, Hermann-von-Helmholtz-Platz 1. 76344 Eggenstein-Leopoldshafen

Abstract

Most of word embedding techniques provide only one vector representation for each word in a text corpus, despite the fact that a single word could have multiple meanings. In this paper, we developed a domain-specific word and phrase embedding model for the patent domain. It treats patent phrases as single information units. Natural language processing techniques are used to extract the meaningful terms from five million patent documents, and a word embedding algorithm is used for generating semantic representation of those terms. This model can be used for a wide rage of tasks like search query expansion, patent semantic similarity search, enrichment and for supporting other patent text mining tasks like patent technology categorization, and knowledge discovery.

1. Introduction

Word embedding techniques have been widely adapted in modern Natural Language Processing (NLP) applications, where words are represented as numerical vectors that capture the semantic and syntactic characteristics of each word in the corpus. Relations with other words as well could be exploited from word embedding models. [1]. Various previous works involved word embedding techniques on patent domain [2] [3]. While the model encodes each word with all of its potential meanings into a single vector, it is unable to model polysemous words which have multiple meanings [4]. For instance, the meaning of the word ”oil” varies with context (e.g; organic, or mineral). The idea behind our phrase embedding is that we want to have different vectors for the different senses. The straightforward solution is just to have two inputs for ”oil”, for example, ”oil_o” and ”oil_m”. Therefore we applied sophisticated NLP tasks to extract the most important patent terms (words and phrases), and then each phrase is merged into a single token, so that it will be represented by a single vector. The remaining of this paper is organized as follows. Section 2 provides our methods for creating the phrase embedding model. Section 3 describes some patent applications that use the phrases embedding model such as word/phrase similarity, patent semantic similarity search, and enrichment tool for patent technology categorization. Finally, Section 4 gives the conclusion and outlines the future research.

2. Construction of Word/Phrase Embedding Model

Word embedding is a technique used to represent words as dense vectors in a vector space. When pre-trained on a large number of tokens, relations of these vectors could make a machine learning system understand the word context and semantics. The key idea is that, because words/phrases with similar meanings often appear in similar contexts, the corresponding embeddings will also be similar. We go beyond a single word embedding to do a better task of modelling complicated concepts. For instance, there is no individual usage of the term ”oil” when it refers to the concept ”vehicle oil”, or ”vegetable oil”, because they have different meanings. Therefore, we create our embedding model not only on single words, but also with phrases. First of all, natural language processing techniques are used to extract the meaningful terms (words or phrases) from patent texts. We then applied the skip-gram of Word2vec algorithm to produce the embedding model.

Published in "Proceedings of The 1st Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2019)", Eds. L. Andersson, H. Aras, F. Piroi, A. Hanbury, PatentSemTech 2019, 12 September 2019, Karlsruhe, Germany. https://doi.org/10.34726/pst2019.8 | © Authors 2019. CC BY 4.0 License.

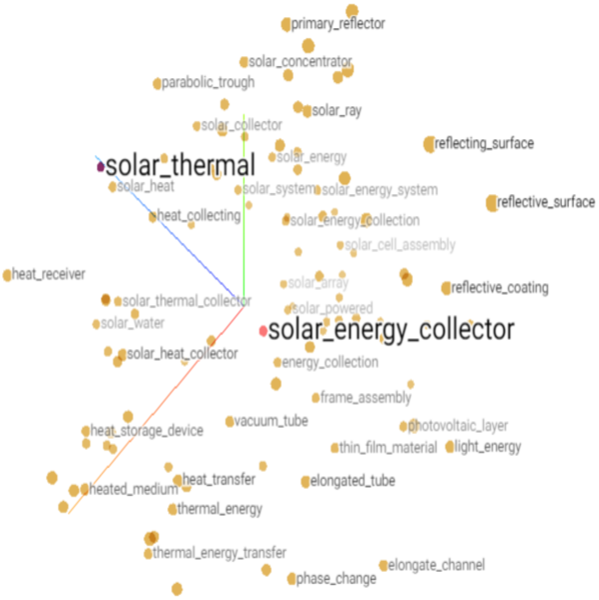


Figure 1: "solar energy collector" space

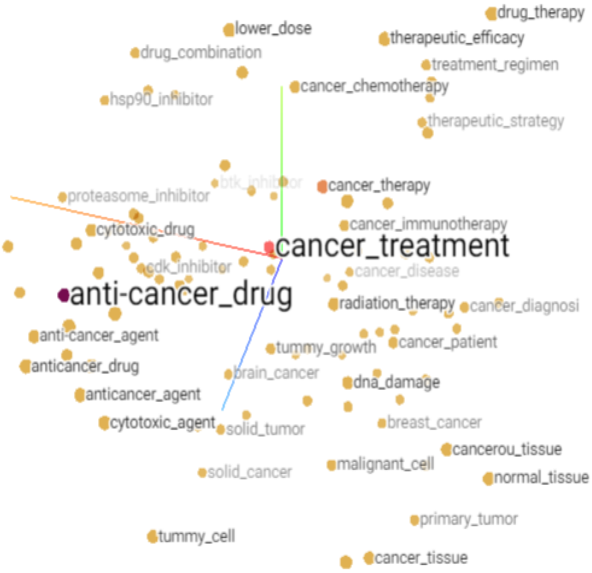


Figure 2: "cancer treatment" space

2.1. Dataset

In this work, we collected and processed approximately five million patents published between 1978 and 2016. These patent documents were extracted from databases of the European Patent Office (EPO) and the World Intellectual Property Organization (WIPO) by using our paten search engine system called STN[2]. From each patent document, we extract the title and abstract that provide the accurate description of the invention.

2.2. Natural Language Processing (NLP)

Many different Natural Language Processing tasks are performed on the provided texts to automatically extract the most significant words and phrases from the collection of patent documents. These NLP tasks include Sentence Detection for transforming the row texts of titles and abstracts into sentences, tokenization for dividing the texts into tokens, part-of-speech tagging and Shallow Syntactic Parsing (or Chunking) for splitting long sentences into smaller chunks. These NLP tasks are applied to transform the collection of title and abstracts into a collection of more than 13 million sentences, including approximately one billion words. Moreover, for each phrase in each sentence, we applied additional NLP tasks: Stop-word Removal, Lemmatization, Pruning (ignoring terms that have very low or very high frequency), and We keep some special characters like "-" that is frequently used (e.g; anti-cancer and multiple-processor). Additionally, a custom n-gram model is used for the long noun phrases. Additional rule-based methods are applied for filtering out undesired words and phrases on the one hand, and for enhancing other phrases on the other hand. For terms that contain more than one word, we linked them via underscores and treated them as single tokens, e.g., "digital_rights_management" and "quantum_cryptography", so that they are represented as single vectors. We implemented our NLP pipeline by using spaCy[3] which is a Python-based library that provides a variety of practical tools for text processing in multiple languages.It is powerful and advanced library used as standard for practical NLP due to its speed, robustness and near state-of-the-art performance.

2.3. Train The Model

After applying our NLP tasks to extract the most significant terms (words and phrases) from the collection of patent documents, we need to create

[2] https://www.stn.org

[3] https://spacy.io/

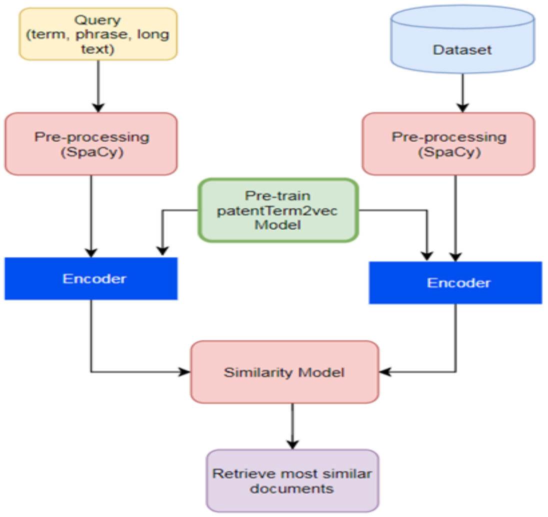

Query (term, phrase, long text)
Dataset
Pre-processing (SpaCy)
Pre-processing (SpaCy)
Pre-train patentTerm2vec Model
Encoder
Encoder
Similarity Model
Retrieve most similar documents

Figure 3: Patent Semantic Similarity Search

vector representations for those terms in an unsupervised manner using a word embedding algorithm. After applying a series of testing experiments, we decided to used the Word2vec implementation in Gensim[4] with few modifications. We experimentally found that skip-gram with negative sampling loss (n = 15) performed best. We used 100-dimensional embeddings, a learning rate equals to 0.01, epochs = 30, window is set to 10, and a minimum word/phrase count is set to 5. The details of this model are available on Github[5].

3. Applications

3.1. Word/Phrase Semantic Relatedness

Generally speaking, semantic relations also enable query expansion to make phrase-related searches more intelligent. Given the fact that word embedding based query expansion methods perform better than co-occurrence based statistics, we can utilize the phrase embedding to expand the query with terms that are semantically related to the query as a whole or to its terms. In addition, this model can be used to find certain word/phrase synonyms. Figure 1 and figure 2 show a dimensional projection of the word embedding space for two phrases ”solar energy collector” and “cancer treatment”, respectively. Additional, our phrase embedding model can capture the relationships between terms and their acronyms, for instance the two technical terms ”Digital Rights Management” and ”Support Vector Machine” are similar to DRM and SVM respectively.

3.2. Patent Semantic Similarity Search

.

We exploited the pre-trained model in order to allow matching queries to patent documents based on meaning rather than keywords. It understands that two terms ”car” and ”vehicle” have similar meanings (synonymy). Instead of building a query with keywords, the user can also describe his/her invention with a few sentences, considering any part of patent texts such as abstract, technical field, summary, or independent claims, in order to get the similar inventions.

Figure 3 illustrates the workflow of the proposed service: for each document in the dataset, we apply the same pre-processing steps that were applied for training the word embedding model in order to get the important terms (words and phrases). Each term in the document is encoded into a vector based on the pre-trained word2vec model. Then, the token vectors are averaged to represent the document into a single vector since each phrase is treated as a single token and hence represented by a single vector. Similarly, we used the same scenario to get the embedding vector of the query. Cosine distance is used to calculate the similarity between the query vector and each document’s vector in the dataset, so that the most similar documents will be returned.

3.3. Patent Technology Categorization

Various methods that involve machine learning and text mining have been proposed in order to extract values from patent information [5] [6]. Patent classification is a kind of knowledge management where documents are assigned into predefined categories. Examples of patent classification systems are the International Patent Classification (IPC) and the Cooperative Patent Classification (CPC), which have hierarchical schemes and used by patent information professionals to assign classes to each patent document. In another work[7], our word/phrase embedding model has been used as transfer learning for patent classification task[6]. Particularly, instead of representing

[4] https://radimrehurek.com/gensim/

[5] https://github.com/sofean-mso/PatentPhraseEmbedding

[6] https://github.com/sofean-mso/DeepL4Patent

Table 1: The top 20 most similar terms to the three patent terms: "quantum computing", "brain tumor", and "autonomous driving"

quantum computing	**brain tumor**	**autonomous driving**
quantum computing device	brain tumor therapy	autonomous driving mode
quantum computer	brain tumor marker	autonomous driving apparatus
qubit device	metastatic brain	autonomous driving control
superconducting flux	brain tumor cells	automated driving
quantum logic	metastatic brain lesions	autonomous driving vehicle
quantum device	human gbm	lane keeping
superconducting phase-charge qubit	brain tumor recurrence	road curvature
hybrid qubit	brain tumors	manual driving mode
quantum circuit	ewing sarcoma	park assist
quantum information processing	primary tumour	target lane
ising model	brain tumor treatment	autonomous driving state
quantum bits	high-grade gliomas	driving assistance
superconducting processor	malignant melanoma	driving environment
quantum processor	brain tumor tissue	driver assistance system
quantum gates	breast tumor	driver interaction
quanton	brain cancer	park assist system
quantum systems	tumor response	travel segment
qubit state	breast cancer	autonomous driving assistance
superconducting integrated circuits	metastatic brain tumor	vehicle surroundings monitoring
quantum interference	malignant brain	automatic travel

each document with a sparse vector, each word in the document can be encoded into n-dimensional vector which called embedding. For example, for a document with two terms "solar collector", and "solar cell", a pre-trained word/phrase embedding model can be used to get embedding of each term and add them together then the document is represented as a single point.

Another patent mining task is an automatic categorization of topic-related patents according to specific categories such as technology, effect, application, use, etc [8] [9]. This task requires high quality training dataset that involves patent experts for manually processing unstructured patent documents. Therefore, phrase embedding model can be used as a tool to help patent experts and data scientists to create a knowledge base for the categorization task since each category can be described by a group of similar/related terms. Table 1 presents the result of retrieving the top 20 most similar terms to the three patent terms: "quantum computing", "brain tumor", and "autonomous driving".

4. Conclusion

In this paper, we developed a domain-specific word and phrase embedding model for the patent domain. The key idea is getting similar embeddings for words and phrases that appear within similar contexts in patent documents. This model can serve many real patent applications such as search query expansion, patent semantic similarity search, enrichment and supporting other text mining tasks like technology categorization. One of our future research directions is going towards developing a knowledge graph for patent terms by using unsupervised methods such as word/phrase embedding.

References

[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean, Distributed representations of words and phrases and their compositionality, Proceedings of the 26th International Conference on Neural Information Processing Systems (2013)

[2] Julian Risch, and Ralf Krestel , Domain-specific word embeddings for patent classification, Data Technologies and Applications, (2019)

[3] Aaron Abood and Dave Feltenberger, Automated patent landscaping, Artif Intell law (2018)

[4] Andrew Trask, Phil Michalak, and John Liu: SENSE2VEC - A fast and accurate method for word sense disambiguation in neural word embeddings. Digital Reasoning Systems, Inc. 2015

[5] Assad Abbas, Limin Zhang, and Samee U.Khan. A literature review on the state-of-the-art in patent analysis, World Patent. Information. 2015

[6] Mustafa Sofean, Hidir Aras, and Ahmad Alrifai. A Workflow-based Large-Scale Patent Mining and Analytics Framework. Proceedings of 24th International Conference on Information and Software Technologies.(ICIST); Vilnius, Litauen. (2018)

[7] Mustafa Sofean. Deep Learning based Pipeline with Multichannel Inputs for Patent Classification. in Proceedings of of PatentSemTech, 1st Workshop on Patent Text Mining and Semantic Technologies (2019)

[8] Anthony Trippe. Guidelines for Preparing Patent Landscape Reports. Patinformatics, LLC, With contributions from WIPO Secretariat. (2015)

[9] World Intellectual Property Report. https://www.wipo.int/patentscope/en/programs/ patent landscapes/plrdb search.jsp?territory code=IN (last access October 2019)