Krestel, Ralf et al.

Conference Paper — Accepted Manuscript (Postprint)

4th Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech2023)

Suggested Citation: Krestel, Ralf et al. (2023) : 4th Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech2023), In: SIGIR 2023, 4th Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech2023), ACM, New York, pp. 3483-3486, https://doi.org/10.1145/3539618.3591929

4th Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech2023)

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

Hidir Aras
hidir.aras@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure
Eggenstein-Leopoldshafen, Germany

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

Florina Piroi
florina.piroi@tuwien.ac.at
TU Wien & Research Studios Austria
Vienna, Austria

Allan Hanbury
allan.hanbury@tuwien.ac.at
TU Wien
Vienna, Austria

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

ABSTRACT

Information retrieval systems for the patent domain have a long history. They can support patent experts in a variety of daily tasks: from analyzing the patent landscape to support experts in the patenting process and large-scale information extraction. Advances in machine learning and natural language processing allow to further automate tasks, such as paragraph retrieval or even patent text generation. Uncovering the potential of semantic technologies for the intellectual property (IP) industry is just getting started. Investigating the use of artificial intelligence methods for the patent domain is therefore not only of academic interest, but also highly relevant for practitioners. Compared to other domains, high quality, semi-structured, annotated data is available in large volumes (a requirement for supervised machine learning models), making training large models easier. On the other hand, domain-specific challenges arise, such as very technical language or legal requirements for patent documents. The focus of the 4th edition of this workshop will be on two-way communication between industry and academia from all areas of information retrieval in particular with the Asian community. We want to bring together novel research results and the latest systems and methods employed by practitioners in the field.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Digital libraries and archives**; **Data mining**; • **Applied computing** → **Document management and text processing**; • **Computing methodologies** → **Natural language processing**.

KEYWORDS

patent analysis, text mining, semantic technology, deep learning

1 MOTIVATION

The PatentSemTech2023 workshop is the forth in a series of workshops that has started in 2019 [1, 9, 10][1]. The aim of the workshop series is to establish a long-term collaboration and a two-way communication channel between industry and academia from relevant fields in information retrieval and related areas in order to explore and transfer new knowledge, methods and technologies for the benefit of industrial applications as well as support interdisciplinary research in applied sciences for the Intellectual Property (IP) and neighbouring domains, such as scientific text analysis from biomedicine, chemistry, etc.

Challenges of using IP data for IR.. Users of patent information systems are highly specialized information professionals and domain experts, who cooperate with research and/or legal departments in their institutions or companies. Finding the right information in the patent databases is critical to business success. There are high requirements for the correctness and completeness of the data to search through, the efficiency of the search interface, the trustworthiness of the provider, and for the quality and completeness of the search results. For general natural language documents (such as news articles, or Wikipedia articles) there are a variety of tools and methods to process and prepare them for a variety of specific task. It is a great challenge to adapt or re-design such tools to address the requirements of working with patent and legal documents.

Patent Data Traits. Patents are a type of scientific text which is complex and difficult to analyse compared to common texts. One reason is that patents, as a corpus and as a single document, are both very heterogeneous. A patent corpus covers very diverse scientific subject areas, such as chemistry, pharmacology, mining, and all areas of engineering, with the consequence that all kinds of terminology can be found in a patent corpus. Further, a patent corpus usually covers a long time span, often from the 1950s to the present. Patents are composed of detailed descriptions of the invention and the patent claims, making them, on average, two to five times longer than scientific articles. Also, patents are usually characterized by the use of legal language to describe what is claimed by the invention, making them hard to understand by laypersons. And finally, typographical errors are not uncommon, since many patents in their machine-readable form are derived from OCR processing and/or machine-translation.

Why work with Patent Data? Patent data, besides its challenging aspects, comes with a richness of facets that makes it interesting for text-mining and semantic methods:

[1] http://ifs.tuwien.ac.at/patentsemtech/

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

Suitability to SIGIR.. Given the traditional connection of patent retrieval and SIGIR [6, 8] on the one hand, and the prevalent use DL methods in information retrieval [11] on the other hand, a dedicated workshop with a focus on DL methods for patent retrieval and analysis will help to bring state-of-the-art IR and AI methods into the patent domain. Further, patent analysis and retrieval are inherently application-driven and IR research is highly relevant for IP practitioners. Within SIGIR, we find the ideal setting to discuss and make connections between IR researchers and practitioners from the IP industry and research at the intersection of AI and patent retrieval. IP practitioners will benefit from the IR community by learning novel IR and AI models that could possibly be applied to patent-related texts, and IR researchers learn about publicly available, well-annotated, large document collections and exciting tasks relevant to industry needs and suitable for IR research.

2 THEME AND MAIN TOPICS

The purpose of this workshop is to motivate researchers in academia and industry to explore, among others,

- text and data mining methods, in particular deep learning based methods,
- semantic enrichment of large amounts of scientific texts, e.g., to aid retrieval systems or generate training data,
- exploit technical information outside of the IP world, for example, by interlinking valuable knowledge sources from domain-specific knowledge graphs (bio-pharma, chemistry, engineering, etc.) or the Linked Open Data (LOD) cloud.

Further, we welcome contributions along our main theme: *Application of Deep Learning Methods for Patent Retrieval and Analysis*, that, among others,

- present novel datasets for training deep learning models,
- explore IP applications with underlying advanced NLP, TDM, and artificial intelligence methods, by applying and adapting DL methods for various domain-specific tasks,
- apply enhanced machine learning and semantic technologies to enrich and analyse patent texts, e.g., in order to contribute to use cases such as technology analysis, trend analysis, semantic patent landscaping, competitor analysis, etc.
- show proof-of-concept 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.

Current research trends in the patent domain show interesting developments in all of these areas. Novel resources area are presented, e.g., patent-specific word embeddings [14] and novel tasks for automation [13]. Efforts in patent landscaping[2] and visualization are also increasing.[3] A literature review resource with the aim to bridge over different scientific text genres is developed by Artificial Researcher IT GmbH [2][4].

3 FORMAT

The program will reflect the main workshop series theme: bridging the gap between patent practitioners and IR researchers by alternating sessions from a research and an industry track. As in previous editions, we encourage national and international patent offices (such as USPTO, WIPO and EPO) and active companies in industry (such as Patsnap, Octimine, Google Patents, InfoChem, etc.) as well as initiatives such as "The Lens" [5] and "ML4Patents.com" [6] to contribute to the workshop as a partner, either by presenting work or attending the discussions.

Anticipated Full-Day Schedule.

- Welcome address (15min)
- Research track session (15+10min per full paper, 10+10min per short paper)
- Keynote (speaker tbd.)
- Patent summarization track session (intro 15min, 5min elevator pitch per short paper + 45min panel discussion with all track participants)
- Systems/Demos track session (10+10min per short paper)
- Open discussion: patents and AI (60min)
- Closing: "I like, I wish…", Feedback by participants (15min)

4 SOLICITED CONTRIBUTIONS

In this year's edition, we solicit two types of submissions: full papers and short papers for three tracks: research, demo, and summarization task. Full papers will be limited to 8 pages; short papers will be 4 pages. The submissions will be reviewed by at least two PC members and selected for presentation based on novelty, interestingness, and impact. We will invite PC members from our previous editions and extend this pool if necessary. The accepted contributions will be published as CEUR proceedings. Selected contributions will be invited to submit extended versions to Elsevier's World Patent Information (WPI) journal[7].

While we still encourage discussion of ideas and not necessarily only presenting mature work, we also want to be an outlet for students to present their results and get feedback from peers. To emphasize the workshop character even more, we explicitly invite experts from industry to present exciting ongoing work and bring together — as in previous editions — academic and industry research in the patent domain. We plan for three tracks:

Research Track. For this track, we solicit contributions from academia that present:

[2] https://github.com/google/patents-public-data
[3] https://wipopearl.wipo.int/en/conceptmap
[4] https://passageretrieval.artificialresearcher.com/
[5] https://www.lens.org
[6] https://www.ml4patents.com
[7] https://www.journals.elsevier.com/world-patent-information

- Novel applications of existing state of the art methods for the IP domain
- Novel methods or tasks in the IP domain
- Novel user interfaces for the IP domain
- Novel evaluation or analysis insights in the IP domain
- Novel benchmark datasets or other resources of interest
- A survey or overview related to a particular task in the IP domain

Given the importance of training data for machine learning research, we especially emphasize the importance of in-depth description of the utilized resources and encourage the open publication of datasets whenever possible.

Systems/Demo Track. We further welcome extended abstracts describing demos, case study, insights, or novel ideas from industry. These contributions should describe a focused case study making use of semantic technologies or machine learning, an interesting IP-related task descriptions or best practices for patent analysis. Demo contributions should describe in-use systems or prototype implementation of semantic technologies or deep learning approaches that can be presented and demonstrated. The focus of the demo can be on processing or analysing data from the IP domain, or focused on user experience or interfaces. We are also very interested in learning about in-use resources related to patents or related legal documents. Further, they can describe external resources to augment IP datasets, e.g., linked open data.

Patent Summarization Track. Within the patent text mining community, especially from the industry, there is an interest in developing text mining tools targeting text summarization, which is not just useful for providing a reader with a quick overview of a patent but has also been successful as the text source for query formulation [3, 7, 12] Participants will be free to use any of the publicly available data sets for training their models. We recommend that participants have a look at prior work [15–17] and to inspect the work and data provided by [4, 5].[8] Participants will submit their solutions which will be evaluated by the workshop organizers. The most interesting submissions will be invited to submit a short paper and present their solution at the workshop.

The participants will be asked to set their solution up as a service and provide a REST API with input: patent document or a scientific document in the format, and output: a summarization in predefined JSON format and the summary shall not be more than 600 tokens (words). We have decided upon the interactive evaluation schema since it provides a more real-life scenario, which makes it possible to evaluate not only the performance in terms of F1, ROUGE, recall, precision etc. but also measure the solution's robustness and response time, which is highly relevant to the industry due to a backlog of millions of documents that need to be summarize as well as a need for interactive solutions.

5 ORGANIZERS

Ralf Krestel is Professor for Information Profiling and Retrieval at ZBW - Leibniz Information Centre for Economics and Kiel University. His research centers around text mining, information retrieval, recommender systems, natural language processing, and machine learning. He has published multiple papers on patent retrieval, e.g., at the recommender systems conference.

Hidir Aras is a senior researcher and head of the applied science department (Patents4Science) at FIZ Karlsruhe, Germany. His research interests include big data analytics, text and data mining, and semantic analysis of patent information. During his time at the University of Bremen, where he finished his doctoral thesis, he focused the use of Semantic Web technologies, social networking paradigms and human computer interaction (HCI) for the interaction and exploration of large information spaces on the Web.

Allan Hanbury is Professor for Data Intelligence and head of the E-Commerce Research Unit in the Faculty of Informatics, TU Wien, Austria. He is also faculty member of the Complexity Science Hub Vienna. He is coordinator of DoSSIER, a Marie Sklodowska-Curie Innovative Training Network, educating 15 doctoral students on domain-specific systems for information extraction and retrieval. Most recently, he was Tutorial Co-Chair of the ECIR 2020 and Short Paper Co-Chair of the ECIR 2018.

Linda Andersson is the CEO of Artificial Researcher-IT GmbH. In 2018 they launched the product idea 'Artificial Researcher in Science' which received the Commercial Viability Award from the Austrian Angel Investors Association. She has over 17 years of experience in the text mining industry and worked with different aspects of scientific literature text mining. She has been working in the Intellectual Property industry, designing different domain-specific patent text mining solutions. Ms Andersson is guest editor for the World Patent Information Issue associated with "Text Mining and Semantic Technologies in the Intellectual Property Domain".

Florina Piroi is a senior scientist at TU Wien, Austria, and at the Research Studios Austria, Data Science group, with experience in domain specific search, search engine evaluation and running evaluation campaigns. She has been coordinating the CLEF-IP evaluation campaign and organising workshops, where specific IR methods for the IP domain have been assessed. Dr. Piroi has been a Lab Chair for CLEF 2021 (Conference and Labs of the Evaluation Forum) and has also been on the Organisation Committee of the ECIR 2015. She is an organizer of the LegalIR[9] workshop that will take place at ECIR 2023.

Dean Alderucci is the director of research for the Center for Artificial Intelligence and Patent Analysis at Carnegie Mellon University. His research involves extracting knowledge from the text of legal and other documents, and automating complex tasks performed by knowledge-intensive workers such as lawyers, regulators, and medical professionals. He also advises organizations on best practices for implementing machine learning and natural language processing technologies, and on creating AI tools customized to various business areas. Dean speaks frequently on applying AI, especially in patent and legal domains. Most recently he was an organizer for the AI and Patent Data workshop at the 33rd International Conference on Legal Knowledge and Information Systems conference and the Learning Network Architecture During Training workshop at the 35th AAAI Conference on Artificial Intelligence.

[8] https://github.com/nlpTRIZ/SummaTRIZ, https://github.com/nlpTRIZ/PaGAN

[9] https://tmr.liacs.nl/legalIR/

6 PREVIOUS EDITIONS OF THE WORKSHOP

1st Edition of the PatentSemTech at SEMANTiCS Conference 2019. The PatentSemTech2019 workshop was the first edition in a planned series of workshops on patent text mining and semantic technologies. Seven papers passed the peer review process (three long, two short and two demo papers). Three submissions that passed the reviewing process were proposed for publication to the World Patent Information (WPI) journal's virtual special issue on "Patent Text Mining and Semantic Technologies". The participants' feedback was positive with the recommendation to continue the good mix of scientific and practical presentations and the demos. The participating experts expressed that such events are too rare, though highly welcomed by both IP experts and academic researchers.

2nd Edition of the PatentSemTech at SIGIR Conference 2021. On July 15th, 2021, the PatentSemTech'21 workshop took place as a one-day online event in conjunction with the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'21).[10] The well-attended event with more than 40 participants from research and industry focused on the application of new machine learning methods for patent retrieval, patent text mining, and patent analysis. In addition to the presentation of scientific papers, various demos and case studies were presented in the workshop — for example, Siemens PatentExplorer or WIPOPearl from the World Intellectual Property Organization. Short presentations on topics related to the application of machine learning methods and semantic technology addressed, for example, Linked Open Data (LOD) in the context of special use cases in industry and publicly funded projects such as IPLodB. In the joint expert panel, scientists and participants from industry discussed the question: "AI and Patent Analysis - Friends or Foes?" The proceedings of the PatentSemTech'21 workshop have been published as Open Access: http://ceur-ws.org/Vol-2909/. In addition, two of the best papers have been proposed for publication in a Virtual Special Issue (VSI) of the World Patent Information (WPI) Journal.

3rd Edition of the PatentSemTech at SIGIR Conference 2022. The third PatentSemTech'2022 workshop was held as a hybrid event within the framework of SIGIR'2022[11]. A total of about 42 participants from industry and academia attended the workshop. Overall 10 extended abstracts of research work (research track) and case studies from industry (industry track) from the total submissions related to information extraction, search, classification and datasets were presented at the workshop. The highlight was the keynote by Jamie Holcombe (CIO USPTO) and a panel on "AI & Patents". Both sessions reiterated the potential for new applications and soluwhen applying AI methods to tents. USPTO intends to actively support the creation of benchmarks for such models. A first effort represents the Competition on "phrase-to-phrase matching" at Kaggle. A paper on a dataset for this was presented by Google in our workshop.

REFERENCES

[1] L. Andersson, H. Aras, F. Piroi, and A. Hanbury (Eds.). 2019. *Proceedings of the 1st Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech).* https://doi.org/10.34726/pst2019

[2] L. Andersson, M. Lupu, Jo. Palotti, A. Hanbury, and A. Rauber. 2016. When is the Time Ripe for Natural Language Processing for Patent Passage Retrieval?. In *CIKM '16 (CIKM '16)*. New York, NY, USA, 1453–1462. https://doi.org/10.1145/2983323.2983858

[3] EKL D'hondt and Suzan Verberne. 2010. CLEF-IP 2010: Prior Art Retrieval using the different sections in patent documents. (2010).

[4] Guillaume Guarino, Ahmed Samet, Amir Nafi, and Denis Cavallucci. 2020. SummaTRIZ : Summarization Networks for Mining Patent Contradiction. In *19th IEEE International Conference on Machine Learning and Applications, ICMLA 2020, Miami, FL, USA, December 14-17, 2020*, M. Arif Wani, Feng Luo, Xiaolin Andy Li, Dejing Dou, and Francesco Bonchi (Eds.). IEEE, 979–986. https://doi.org/10.1109/ICMLA51294.2020.00159

[5] G. Guarino, A. Samet, A. Nafi, and D. Cavallucci. 2021. PaGAN: Generative Adversarial Network for Patent understanding. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE Computer Society, Los Alamitos, CA, USA, 1084–1089. https://doi.org/10.1109/ICDM51629.2021.00126

[6] A. Hanbury, V. Zenz, and H. Berger. 2010. 1st international workshop on advances in patent information retrieval (AsPIRe'10). *SIGIR Forum* 44, 1 (2010), 19–22. https://doi.org/10.1145/1842890.1842893

[7] Fidelia IbekweSanjuan, Silvia Fernandez, Eric SanJuan, and Eric Charton. 2011. Annotation of Scientific Summaries for Information Retrieval. *CoRR* abs/1110.5722 (2011). arXiv:1110.5722 http://arxiv.org/abs/1110.5722

[8] Noriko Kando and Mun-Kew Leong. 2000. Workshop on Patent Retrieval (SIGIR 2000 Workshop Report). *SIGIR Forum* 34, 1 (2000), 28–30. https://doi.org/10.1145/373593.373621

[9] R. Krestel, H. Aras, L. Andersson, F. Piroi, A. Hanbury, and D. Alderucci. 2021. 2nd PatentSemTech Workshop. In *SIGIR '21: The 44th ACM SIGIR Conference*. ACM, New York, USA, 2693–2696. https://doi.org/10.1145/3404835.3462816

[10] R. Krestel, H. Aras, L. Andersson, F. Piroi, A. Hanbury, and D. Alderucci. 2022. 3rd PatentSemTech Workshop). In *SIGIR '22: The 45th ACM SIGIR Conference*. ACM, New York, USA, 3474–3477. https://doi.org/10.1145/3477495.3531702

[11] Hang Li and Zhengdong Lu. 2016. Deep learning for information retrieval. In *Proc. of the 39th ACM SIGIR conference*. ACM, New York, USA, 1203–1206.

[12] P. Mahdabi, L. Andersson, A. Hanbury, and F. Crestani. 2011. Report on the CLEF-IP 2011 Experiments: Exploring Patent Summarization. In *In CLEF (Notebook Papers/LABs/Workshop)*.

[13] Julian Risch, Nicolas Alder, Christoph Hewel, and Ralf Krestel. 2020. PatentMatch: A Dataset for Matching Patent Claims & Prior Art. *CoRR* abs/2012.13919 (2020), 5 pages. arXiv:2012.13919 https://arxiv.org/abs/2012.13919

[14] J. Risch and R. Krestel. 2019. Domain-specific word embeddings for patent classification. *Data Technol. Appl.* 53, 1 (2019), 108–122. https://doi.org/10.1108/DTA-01-2019-0002

[15] A. Trappey, C. Trappey, and C-Y. Wu. 2008. A semantic based approach for automatic patent document summarization. In *Collaborative Product and Service Life Cycle Management for a Sustainable World*. Springer, 485–494.

[16] A. Trappey, C. Trappey, and C-Y. Wu. 2009. Automatic patent document summarization for collaborative knowledge systems and services. *Journal of Systems Science and Systems Engineering* 18, 1 (2009), 71–94.

[17] A. Trappey, C. Trappey, J-L. Wu, and J. Wang. 2020. Intelligent compilation of patent summaries using machine learning and natural language processing techniques. *Advanced Engineering Informatics* 43 (2020), 101027.

[10] http://ifs.tuwien.ac.at/patentsemtech/2021/

[11] https://sigir.org/sigir2022/program/workshops/