
Ralf Krestel  
rkr@informatik.uni-kiel.de  
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  
daldereu@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, question answering (QA) 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. With the 5th edition of this workshop we will provide a platform for researchers and industry to learn about novel and emerging technologies for semantic patent retrieval and big analytics employing sophisticated methods ranging from patent text mining, domain-specific information retrieval to large language models targeting next generation applications and use cases for the IP and related domains.

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

ACM Reference Format:

1 MOTIVATION

The PatentSemTech2024 workshop is the fifth in a series of workshops that has started in 2019 [1, 4, 5]. 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:

- 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.

Why PatentSemTech at SIGIR? Given the traditional connection of patent retrieval and SIGIR [2, 3] on the one hand, and the prevalent use DL methods in information retrieval [6] on the other hand, a dedicated workshop with a focus on DL methods for patent retrieval and analysis helps 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 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, patent monitoring and recommendation, 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 [8] and novel tasks for automation [7]. Efforts in patent landscaping2 and visualization are also increasing.3

3 SOLICITED CONTRIBUTIONS AND SELECTION PROCESS

In this year’s edition, we solicited two types of submissions: full papers and short papers for two tracks: research and demo. Full papers were limited to 8 pages and short papers were limited to 4 pages. The submissions were reviewed by at least two PC members and selected for presentation based on novelty, interestingness, and impact. 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) journal4.

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 invited 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 two tracks:

Research Track. For this track, we solicited research contributions (short and full-papers) in the IP domain:

- Novel applications of AI methods
- Novel methods or tasks
- Novel user interfaces
- Novel evaluation or analysis insights
- Novel benchmark datasets or other resources
- Surveys or overview papers

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.

References:

1. https://www.journals.elsevier.com/world-patent-information
3. https://www.journals.elsevier.com/world-patent-information

2. https://www.journals.elsevier.com/world-patent-information
**Systems/Demo Track.** We further welcomed short papers 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.

### 4 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, semantic analysis and retrieval, and novel (visual) retrieval interfaces for patents and scientific information. The focus of his research team is on exploring new technologies for the acquisition, linking and use of patent knowledge for the benefit of science and research.

**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 Skłodowska-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. Since 2019 she her team have been designing different domain-specific text mining solutions for the IP and Open Science community.

**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.

**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. In the past, he was an organizer for several other workshops on AI, patents and legal data.

### 5 PREVIOUS EDITIONS OF THE WORKSHOP

This is the fifth edition of the PatentSemTech workshop. We constantly try to improve the workshop format together with the participants, who we encourage to make suggestions and voice their needs and wishes for future editions.

**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 submissions 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). 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. A total of 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 solutions when applying AI methods to patents. 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.

---

3http://ifs.tuwien.ac.at/patentsemtech/2021/
4th Edition of the PatentSemTech at SIGIR Conference 2023. The third PatentSemTech’2023 workshop was held as a hybrid event within the framework of SIGIR’2023. A total of 20 participants from industry and academia attended the workshop. The findings from interactive summarization session highlighted the significant time and effort required to create authentic reference data (ground truth) and the acquired substantial amounts of annotated data essential for training various NLP applications. In this session, participants evaluated three automatically generated patent summaries alongside one manually crafted summary for five distinct patents. To simplify the evaluation process, patents from the text mining domain were chosen; however, the assessment still required approximately 15 minutes per patent from each participant. Hence, the team at Artificial Research GmbH undertook the task of creating a dataset to propel the progress of domain-specific NLP. Even in the absence of fully manually annotated data, their aim was to develop a semi-annotated dataset to support research advancements utilizing DL and NLP to generate training data.

REFERENCES