Diving into Knowledge Graphs for Patents: Open Challenges and Benefits

Danilo Dessí[1,*], Rima Dessí[2]

[1] *GESIS Leibniz Institute for the Social Sciences, Cologne, Germany*
[2] *Patent for Science Department, FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Karlsruhe, Germany*

Abstract

Textual documents are the means of sharing information and preserving knowledge for a large variety of domains. The patent domain is also using such a paradigm which is becoming difficult to maintain and is limiting the potentialities of using advanced AI systems for domain analysis. To overcome this issue, it is more and more frequent to find approaches to transform textual representations into Knowledge Graphs (KGs). In this position paper, we discuss KGs within the patent domain, present its challenges, and envision the benefits of such technologies for this domain. In addition, this paper provides insights of such KGs by reproducing an existing pipeline to create KGs and applying it to patents in the computer science domain.

Keywords

Patent Domain, Knowledge Graph, Intellectual Property

1. Introduction

Publishing patents in natural language textual documents is the current paradigm employed to describe and disclose technological innovations and protect intellectual properties. Searching, analyzing, and understanding patents are keys for analyzing the current state of industry and society, identifying their needs, and driving their future development. However, these key points may not be achieved with today's publishing paradigm due to the complexity, heterogeneity, and length of patent documents. This limits patent searchers in finding and exploring patents to support business-critical decisions and makes such processes expensive and time-consuming. The main limitation of such a paradigm is given by the typical complexity of the natural language which requires critical human thinking based on grammar, semantics, and a complex understanding of what is conveyed in patents' text [1]. This is also more and more stressed by the striking growth of patents' number which makes the evaluation process of new patents difficult for patent offices [2]. Patent writers as well as other involved stakeholders (e.g., patent offices, academic institutions, industries, and policymakers) usually rely on search engines such as Google Patents[1] which only allow keyword search to get a list o patents related to

SemTech4STLD '23: First International Workshop on Semantic Technologies for Scientific, Technical and Legal Data, May 28th, 2023, Hersonissos, Greece

*Corresponding author.
danilo.dessi@gesis.org (D. Dessí); rima.dessi@fiz-karlsruhe.de (R. Dessí)
0000-0003-3843-3285 (D. Dessí); 0000-0001-8332-7241 (R. Dessí)

[1] https://www.google.com/?tbm=pts

the desired topic, and do not provide any ready-to-use insight about the protected invention. Furthermore, we are today witnessing the birth of new intelligent systems such as ChatGPT which can provide general knowledge about a large variety of domains. However, it is still difficult to rely on such systems due to the fact that they might provide incorrect information, partially invented answers and not verifiable information due to a lack of transparency in the used technology and algorithms [3]. Therefore, there is still a need to explore ad-hoc solutions for sensitive domains which involve technical and legal aspects where intelligent systems can support involved stakeholders to make sense of patent content for a variety of applications.

One prominent solution that is taking place in various domains is the development of Knowledge Graphs (KG), i.e., interlinked graphs of entities that describe a domain based on well-defined and formal semantics, to support a variety of tasks. KGs have been extensively utilized in a variety of domains, including artificial intelligence, semantic web, information retrieval, etc. By structuring facts in a graph-like architecture, KGs enable machines to reason and infer implicit knowledge, and perform sophisticated analysis. Furthermore, KGs draw great attention from researchers, especially after the announcement of Google's Knowledge Graph [4]. Several papers have been published for the creation, completion, and alignment of KGs [5, 6, 7, 8, 9, 10]. Examples of KGs, among others, can be found in the biology [11], scholarly [12, 5], and medical domain [10]. Biology benefits from such KGs because of the fast sharing of new discoveries among several institutions, the scholarly domains obtained benefits for the search and analysis of research trends, medical domain used KGs for clinical decision support systems.

However, there has been comparatively less investment and focus on patent-related KGs creation although first investigations can be found. For example, by utilizing patent claims, authors in [13] aim to create an engineering knowledge graph. They extract fact based on pre-defined rules which are solely related to the engineering domain. The KG could be exploited only for the extraction of technical elements, i.e., engineering knowledge. One drawback of this existing solution is that defining these rules is quite expensive and time-consuming. Similarly to authors in [13], [14] aims to build a patent-KG that consists of facts related to engineering design. The authors use patents that have specific Cooperative Patent Classification (CPC)[2] codes, and thus the contained facts are very domain dependent. As a result, the application of such KG is quite limited to the specific domain. Pipelines and methods for other domains or applicable to a broader range of domains are more and more demanded today.

With this in mind, in this position paper, we sketch the potential challenges of building KGs about patents, we discuss the positive implications that such technologies can have for the stakeholders, and finally, we provide insights into how pipelines for this field might be built by reproducing an existing one. More precisely, the contributions of this paper are:

- We provide an overview of the current state of development of KGs for patents.
- We discuss the challenges of building such KGs for the patent domain, highlighting what are the differences from existing solutions available in other domains.
- We describe how the patent domain can benefit from the development of such resources.
- We introduce our first efforts and results about the use of existing solutions to build knowledge graphs about patents.

[2] https://www.epo.org/searching-for-patents/helpful-resources/first-time-here/classification/cpc.html

2. Challenges in Patent Knowledge Graph Construction

This section describes the challenges around the KGs construction for patents from the complexity of patents' structure, natural language, and use perspectives.

2.1. Complexity of Patents' Structure

Patents pose a significant challenge to process and explore due to their length, structure, and domain-specific vocabulary. Patents are published and preserved through an article-centric paradigm which usually includes a title, an abstract, innovative claims, images, and a detailed description of the invention that is protected. The title is a concise text which introduces the main subject of the patent itself. It differs from titles given to documents of different natures because its purpose is not to raise the interest of the reader; it must precisely describe the item or intellectual property the patent document describes. The abstract is the section of patent documents that in a brief paragraph gives an overview of the content of the patent without exposing too many details about the intellectual property; it is a key element used today by patent writers and patent offices for exploring the patent landscape because of its lightweight complexity to have a first notion of the protected item. Claims are short paragraphs made by only a few sentences that state what patents legally protect and compose the most sensitive section of patents; if an innovative element is not explicitly defined in the claims, other intellectual properties can claim it as an innovation in their respective patent documents. The description section provides detailed information about what a patent protects, lists all its components, and provides specific information about its intended uses. The images section provides pictures that support the description of what is protected and are referenced in the text. Last but not least, patents refer to other patents or scientific publications by listing them as the final section of their document. This variety of sections makes it challenging to convey the intended purposes into KGs. For example, representing in a graph form the relationship that occurs between an image and the text that describes it is not easy. In fact, the process that needs to be semantically represented in the patent should reflect the human behavior of reading the text and connect what is read with the visual information delivered by the image itself. However, such a challenge still remains open. Another challenge related to the structure of patents is the representation of what is protected by the patent claims. This is because a KG describing such part of patent documents should be highly precise and should not contain errors due to the legal requirements; however, this cannot be guaranteed with today's technology and future research is required. Last but not least, the current systems might not be fully precise, the extraction of details might fail, thus creating incomplete KGs which cannot be completed with the currently existing techniques.

2.2. Natural Language Challenges and Limitations

One main challenge given by patent documents is that they store and preserve most of the information in natural language text. Therefore, patent documents inherit challenges and drawbacks typically recognized by the Natural Language Processing (NLP) community. Natural language is unstructured thus it is difficult for machines and automatic systems to parse and use them. Specifically, in the context of KG construction, NLP tools should be employed to extract entities and identify relationships among them. However, this presents several challenges: i) it is not easy to understand whether a text span represents a relevant entity to describe the subject of the patent, ii) the same entity can appear in different shapes (e.g., different texts can be used to refer to the same thing), iii) the same text can refer to two or more different entities iv) natural language is complex and it is difficult by means of triples to reproduce the relationships among entities described in the text. Addressing these challenges is crucial for the patent domain. In fact, if two patents refer to different items but they use the same vocabulary (i.e., the word *cup* used to describe a small bowl-shaped container for drinking[3] is different from the word *cup* used to describe trophies[4]), this must be taken into account while building the knowledge graph by solving tasks such as entity disambiguation. Another important aspect that we would like to highlight is the ambiguity of the natural language which might make misleading or incomplete the information extracted by NLP tools. More precisely, triples in the form <head, predicate, tail> might not contain sufficient context or complete information to be used. For example, given the sentence *A method of assembling a water bottle cap system positioning the small ring around a base of the small cap* from the patent *US9771189B2*[5] the triple `<water bottle cap system, position, small ring>` might be extracted; however, this triple is not fully complete since it does not provide the full context of where the *small ring* should be positioned.

2.3. Knowledge Graph Intended Use

One of the most difficult challenges around the construction of KGs is the definition of usage borders. In fact, several KGs about patents can be envisioned to answer the domain demands. More precisely, the following KGs might be built on the basis of the intended use:

- **Metadata KGs.** This type of KG describes the metadata of patents and represents relationships that exist among patents themselves by representing the patent citation network, relationships that occur among authors and patents, relationships that exist among patent authors themselves, and relationships that occur among patent authors and patent offices. They can also be built by exploiting the keywords associated with patents, or the CPC codes for their classification. Building these KGs requires the definition of proper schemas and ontologies since their main goal is to allow efficient and fast search in a large number of patent documents.
- **Entity Mention KGs.** This type of KGs might be used to describe entities that are directly mentioned in patent documents. More precisely, they can be used to describe whether an entity is specifically created by a patent document, whether an entity is used to create a new innovation, whether an entity is a legally protected item, and so on. Such kinds of KGs present challenges in detecting the entities from the textual or visual content of the patent document. Furthermore, they are also challenging due to the fact that they should specify the role of the identified entity in the KG.

[3] https://patents.google.com/patent/US8807371B2/en?q=(cup+drinking)&oq=cup+drinking

[4] https://patents.google.com/patent/US6783255B1/en?q=(cup+trophy)&oq=cup+trophy

[5] https://patents.google.com/patent/US9771189B2/en?q=(water+bottle)&oq=water+bottle

- **Content-based KG.** These KGs describe the content of patents by extracting and formally representing the relationships between extracted entities by converting the meaning of natural language sentences into RDF triples. These KGs are difficult to be built because of (i) the complexity of formally representing the natural language, and (ii) the kind of intent that should be conveyed in the triples based on the section in which the represented content is placed.

3. Patent Knowledge Graph Benefits

This section describes some of the main benefits that KGs can bring to the patent domain to address patent domain challenges. In fact, due to the rapid growth of online available patent data, efficient and effective analysis of such documents has become a crucial task. Furthermore, it is quite critical for patent searchers to find the right information that can be used to support business-critical decisions.

Efficient Search and Discovery. Patent KGs which illustrate the semantic relations between patents have the potential to facilitate the efficient and effective discovery, exploration, and inference of patent-relevant information. Various information retrieval systems can easily exploit such KGs. For instance, patent landscaping systems that aim to identify patents related to a specific topic, often utilize quite sophisticated models. However, incorporating patent KGs can lead to more efficient and effective solutions. Another example is a patent question-answering system. There has been a lack of effort to develop such systems for patents to this date. However, KGs would provide the basis for researchers to develop a question-answering system for patents.

Provenance. KGs might play a relevant role in tracking novel intellectual properties over time, allowing a deep analysis of what is being developed at specific points in time and, therefore, enabling the preservation of relevant historical facts about patents. For example, KGs can be used to represent the history of patents from the application to the publication, thus enabling the formal representation of the provenance information about new intellectual properties.

Explainability and Interpretability. KGs are becoming more and more important to allow the explainability and interpretability of models applied to any kind of data. The patent domain is not an exception and patent KGs can be relevant to explain why a certain patent is classified under a certain category as well as why a patent document refers to another one. Furthermore, interpretability and explainability assessment criteria (e.g., reliability, causality) are more and more required from society and industry and patent KGs would support such evaluations. Thay would enhance our understanding of patent processes, uncover patterns used by inner mechanisms, and empower patent platforms with systems to increase people's trustworthiness in intelligent systems for patents.

Automatization. KGs can unlock the understanding of how new intellectual properties and items might be exploited for uses they were not designed for. More precisely, patent KGs can be used with state-of-the-art technologies to represent inventions and their characteristics in complex vector models that can be used to find similarities among intellectual properties as well as enable the use of machine learning models on such data for tasks such as classification and clustering.

4. Computer Science Patent KG

In this section, we introduce the reader to our first efforts to build a KG about the content of patent documents. For such purpose, we explain the modules that we are currently experimenting with using a subset of patents about the computer science domain. For this, we reproduce the pipeline described in [5], describe the benefits that a full-fledged KG may bring, and outline future developments tailored to the patent domain.

4.1. Pipeline outlook and Use Case

The used pipeline is based on supervised and unsupervised components. More precisely, it is composed by:

- **Extractor Modules.** This module uses state-of-the-art extractors to find computer science entities from the natural language text. More precisely the used tools are DyGIEpp, the CSO classifier, and the Stanford Core NLP suite. DyGIEpp and the CSO classifier are used to extract entities and a set of predefined relationships. Entities are associated with 5 different types: method, task, material, metric, and other entity. Stanford Core NLP is used to extract verbs that put into relation entities directly from the text. These modules provide the basic set of triples that are used to build the KG.
- **Cleaning Modules.** The cleaning modules use a set of heuristics to lemmatize the entities, merge similar entities, and link the entities to external knowledge bases such as DBpedia and Wikidata. Moreover, this module is also used to map verbs with similar meanings to unique representatives given by a hand-crafted taxonomy (for example, the verbs *use*, *employ*, *utilize*, and *exploit* are all mapped to the same verb use).
- **Classification Module.** The classification module uses transformers to automatically validate the triples which might contain incorrect or misleading triples. For doing so, this module exploits a classifier trained on scientific documents and finetuned on trustworthy triples (i.e., triples that have been frequently extracted and, hence, which have several scientific papers that support their contained information).
- **Ontology-based Module.** This module uses a formally defined ontology for representing the relationships among methods, metrics, tasks, and materials; it maps all the generated triples to such ontology and discards triples that do not comply with its defined semantics.

The pipeline has been applied on 1085 patents published in the years 2017 and 2018 from the *The Harvard USPTO Patent Dataset (HUPD)*[6]. To limit our investigation to the computer science domain, we selected patents whose USPTO class is *Data Processing - Artificial Intelligence*[7]. The generated KG includes more than 3K entities and more than 4K triples. The reader can find the used patents as well as the generated KG at https://github.com/danilo-dessi/patent.

4.2. PatentKG Analysis

In this section, we describe the benefits that can be envisioned with the generated KG. To start with, it allows the investigation of fine-grain patent elements and their interaction. For example,

[6] https://huggingface.co/datasets/HUPD/hupd

[7] https://www.uspto.gov/web/patents/classification/uspc706/defs706.htm

we can observe generated triples describing language generator systems for specific domains e.g., `<query generator`, `uses`, `domain specific language>` , or triples such as `<boltzmann machine`, `analyzes`, `gibbs distribution>` describing complex relationships which have been explored in domains like physics and cognitive sciences. Additionally, such KGs can also provide information about broader concepts that exist within the computer science domain. For example, it is possible to study triples that have been found in more than one patent and, therefore, describe pieces of knowledge that are more common in the domain. Examples of such pieces of knowledge are: `<computer program`, `uses`, `computer storage medium>`, `<ai model`, `uses`, `learner module>`, `<neural network`, `includes`, `synapsis>`. In addition, this kind of KGs can help patent offices and stakeholders to explore how elements have been used in already protected inventions, thus supporting the evaluation process for newly submitted patents. Last but not least, these KGs might enhance the study of the patent domain dynamics by analyzing how protected inventions and their components are related over time, thus providing a means to make sense of the patent landscape.

4.3. Limitation and Development Plan

To better fit the patent domain and overcome some limitations, we plan to revise some modules of the pipeline. To start with, we are developing ad-hoc modules to extract entities and relations for patents. This will allow the new pipeline to focus only on entities that are relevant for patent documents. For doing so, we are experimenting with deep learning models for key-phrase extraction. Second, we will revise the current verb taxonomy to better represent the use of verbs in the resulting knowledge graph. This is in fact an important factor in the patent KG because of the legal nature of its content which restricts the usage and meaning of the vocabulary. Third, we plan to create specific modules for each patent section; as explained in section 3.1, patent sections have a specific intent and thus the ontology used to represent the information of such section should be designed accordingly.

5. Conclusion

In this paper, we presented an overview of the use of KGs for the patent domain. In particular, we highlighted three challenges related to the current paradigm for the protection of intellectual properties. Then, we presented which benefits such technologies can bring in the domain, and envision their use for a multitude of tasks. Finally, we introduce the reader to a reproducibility study by adapting an existing pipeline that is tailored to the scholarly domain to be applied to the patent domain. We present some examples of the befits KG-based technologies can bring to the patent domain.

References

[1] M. Y. Jaradeh, A. Oelen, K. E. Farfar, et. al., Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge, in: Proceedings of the 10th International Conference on Knowledge Capture, 2019, pp. 243–246.

[2] A. Manukyan, D. Korobkin, S. Fomenkov, S. Kolesnikov, Semantic patent analysis with amazon web services, in: Journal of Physics: Conference Series, volume 2060, IOP Publishing, 2021, p. 012025.

[3] A. Meloni, S. Angioni, A. Salatino, F. Osborne, D. Reforgiato Recupero, E. Motta, Integrating conversational agents and knowledge graphs within the scholarly domain, IEEE Access 11 (2023) 22468–22489. doi:10.1109/ACCESS.2023.3253388.

[4] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, et. al., Knowledge graphs, CoRR abs/2003.02320 (2020). URL: https://arxiv.org/abs/2003.02320. arXiv:2003.02320.

[5] D. Dessí, F. Osborne, D. R. Recupero, D. Buscaldi, E. Motta, Scicero: A deep learning and nlp approach for generating scientific knowledge graphs in the computer science domain, Knowledge-Based Systems 258 (2022) 109945.

[6] A. Borrego, D. Dessì, I. Hernández, F. Osborne, D. Reforgiato Recupero, D. Ruiz, D. Buscaldi, E. Motta, Completing scientific facts in knowledge graphs of research concepts, IEEE Access 10 (2022) 125867–125880. doi:10.1109/ACCESS.2022.3220241.

[7] M. Y. Jaradeh, A. Oelen, M. Prinz, M. Stocker, S. Auer, Open research knowledge graph: A system walkthrough, in: A. Doucet, A. Isaac, K. Golub, T. Aalberg, A. Jatowt (Eds.), Digital Libraries for Open Knowledge - 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings, volume 11799 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 348–351. URL: https://doi.org/10.1007/978-3-030-30760-8_31. doi:10.1007/978-3-030-30760-8_31.

[8] M. Y. Jaradeh, A. Oelen, K. E. Farfar, M. Prinz, et. al., Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge, in: M. Kejriwal, P. A. Szekely, R. Troncy (Eds.), Proceedings of the 10th International Conference on Knowledge Capture, K-CAP 2019, Marina Del Rey, CA, USA, November 19-21, 2019, ACM, 2019, pp. 243–246. URL: https://doi.org/10.1145/3360901.3364435. doi:10.1145/3360901.3364435.

[9] J. Martínez-Rodríguez, I. López-Arévalo, A. B. Ríos-Alvarado, Openie-based approach for knowledge graph construction from text, Expert Syst. Appl. 113 (2018) 339–355. URL: https://doi.org/10.1016/j.eswa.2018.07.017. doi:10.1016/j.eswa.2018.07.017.

[10] L. Li, P. Wang, J. Yan, Y. Wang, S. Li, J. Jiang, Z. Sun, B. Tang, T. Chang, S. Wang, Y. Liu, Real-world data medical knowledge graph: construction and applications, Artif. Intell. Medicine 103 (2020) 101817. URL: https://doi.org/10.1016/j.artmed.2020.101817. doi:10.1016/j.artmed.2020.101817.

[11] L. Amos, D. Anderson, S. Brody, A. Ripple, B. L. Humphreys, Umls users and uses: a current overview, Journal of the American Medical Informatics Association 27 (2020) 1606–1611.

[12] D. Dessí, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta, Cs-kg: A large-scale knowledge graph of research entities and claims in computer science, in: The Semantic Web–ISWC 2022: 21st International Semantic Web Conference, Virtual Event, October 23–27, 2022, Proceedings, Springer, 2022, pp. 678–696.

[13] L. Siddharth, L. T. Blessing, K. L. Wood, J. Luo, Engineering knowledge graph from patent database, Journal of Computing and Information Science in Engineering 22 (2022) 021008.

[14] H. Zuo, Y. Yin, P. Childs, Patent-kg: Patent knowledge graph extraction for engineering design, Proceedings of the Design Society (2022).