Annotating Discursive Roles of Sentences in Patent Descriptions

Lufei Liu, Xu Sun, François Veltz, Kim Gerdes

Annotating Discursive Roles of Sentences in Patent Descriptions

Lufei Liu[1] and **Xu Sun**[1,2] and **François Veltz**[1] and **Kim Gerdes**[1,3]
[1]Qatent, Paris, France
[2]Université Paris Cité, France
[3]Université Paris-Saclay, Lisn (CNRS), France
{lufei, francois, kim}@qatent.com, xu.sun@etu.u-paris.fr

Abstract

Patent descriptions are a crucial component of patent applications, as they are key to understanding the invention and play a significant role in securing patent grants. While discursive analyses have been undertaken for scientific articles, they have not been as thoroughly explored for patent descriptions, despite the increasing importance of Intellectual Property and the constant rise of the number of patent applications. In this study, we propose an annotation scheme containing 16 classes that allows categorizing each sentence in patent descriptions according to their discursive roles. We publish an experimental human-annotated corpus of 16 patent descriptions and analyze challenges that may be encountered in such work. This work can be base for an automated annotation and thus contribute to enriching linguistic resources in the patent domain.

1 Introduction

Patent applications represent a first step in obtaining exclusive rights over an invention. Analyzing these documents enables inventors to understand technological trends, avoid potential litigation, and assess the competition. The *patent description*, a substantial part of the patent application, provides detailed information about the invention. Although specific segments have to be present in order to have the application accepted, the information can be provided without any pre-imposed order. Patent descriptions should be well organized in order to communicate the invention's technical details, advantages, and scope with more clarity. This, in turn, helps patent examiners to review applications more efficiently, reduces the likelihood of misinterpretation or ambiguity, and increases the chances of obtaining a patent grant with a well-defined scope of protection.

The contributions of our work are as follows:

- Introducing an annotation scheme based on and adapted for the discursive structure of patent descriptions.
- By focusing on patent descriptions, we aim to contribute to a better understanding of these documents' linguistic characteristics and structures, which have received little attention in patent-related research.
- Set a ground for future patent description analysis, for example, develop automatic methods to apply the annotation to large scale patent datasets. This can contribute to the detection of abnormal patent description and study the patent writing style according to different assignees.

2 Related work

The analysis of document structure allows for a deeper understanding of the author's thought process and facilitates the retrieval of specific information within the document. Many previous studies have focused on the analysis of technical documents, particularly scientific papers. For example, (Fisas et al., 2015) created a multi-layered annotated corpus of 40 scientific papers in the domain of Computer Graphics, with each sentence annotated according to its rhetorical role. (Dasigi et al., 2017) created a corpus by manually annotating 75 articles in the domain of intercellular cancer pathways. Each article was divided into clauses which are classified into one of the following categories: Goal, Fact, Result, Hypothesis, Method, Problem, Implication, None. This corpus was used to develop a discourse tagger for claim extraction and evidence fragment detection (Li et al., 2021).

Patent applications are another type of technical document that has garnered researchers' interest. A patent application usually consists of various components, including a title, abstract, description, one or more claims, drawings, and classification information. Current patent text analysis mainly focuses

Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII), pages 235–243

on claims or abstracts to improve claim readability (Ferraro et al., 2014; Okamoto et al., 2017), such as using them to build an engineering knowledge graph (Siddharth et al., 2021; Zuo et al., 2022), or to aid in patent classification models (Lee and Hsiang, 2020). However, the less-structured and much longer patent descriptions, an essential part of understanding patents, receive little attention. To our knowledge, only (Nakamitsu et al., 2022) analyzes the structure of patent descriptions, but they focus solely on four content types: Field, Problem, Solution, and Effect. Nevertheless, patent descriptions contain much richer information, including technical term definitions in context, advantages of the invention, and drawing descriptions. Exploring the structure of patent descriptions can be used to acquire patent writing skills – for humans and machines. Writing a good patent description not only requires an understanding of legal knowledge, but also requires expertise in relevant technical fields. Furthermore, mastering the structure of a patent description enables the extraction of reliable features, which may be useful for patent text modeling, specific to domains, assignees, and legal goals of the patent.

The goal of this study is to apply the discourse structure analysis, a common practice in scientific papers, to the whole patent descriptions while considering their unique writing style. To achieve this, we design an annotation scheme to label each sentence in the description according to its discursive role.

3 Annotation scheme

The patent description is typically divided into several sections. Under the Patent Cooperation Treaty (PCT), the description contains mainly (WIPO, 2022): *Title of invention*, *Technical field*, *Background art*, *Summary of invention*, *Brief description of drawings*, and *Description of embodiments*. The field section specifies the **technical domain** to which the invention belongs. The background section discusses the **prior art** related to the invention, identifies previously encountered **problems**, and explains how the proposed invention may offer solutions to one or more of these issues. The summary section highlights the **key features** and **advantages** of the invention. This is commonly followed by a section that provides a concise overview of the content present in each illustrative drawing, if they are included. Lastly, the Detailed Description section should encompass greater detail of the **claimed** invention by way of **examples (embodiments)**, **describing figures** in detail and **defining** little known or specially formulated technical terms when necessary, to further clarify the structure and functioning of the invention.

Based on the above essential elements recognized in a patent description and with the assistance of a patent attorney, we initially designed a set of 12 labels corresponding to the bold elements above. **key features** and **examples (embodiments)** were combined, represented by the label *Embodiment*, as the invention is usually described by introducing its features. Following this, two annotators collaboratively annotated two patent descriptions and identified a need to distinguish between the *Advantage* and *Problem* labels, to clarify whether these pertain to the invention itself or to existing technologies. In addition, we added the label *Other* for sentences that don't fall into any of the established 14 categories.

This set of 15 labels was applied by two annotators on the first test dataset of 8 patent applications. We noticed that the *Section title* label does not cover all kinds of titles within a patent description, since different applicants may introduce subsections with additional titles according to their writing styles. The annotators found it difficult to decide on the class of non-standard titles: Are they section titles or part of the embodiment?

To remedy this difficulty, a 16$^{\text{th}}$ label *Section subtitle* was added following the annotation of the first dataset. This new label also allows for an elementary encoding of the scope of the main sections, whenever they are indicated by section titles. It is this set of 16 labels that has reached consensus and is deemed operational and representative for annotating the discursive role of sentences within patent descriptions.

3.1 Annotation tags

Below is a brief summary of the labels defined for annotating patent descriptions. Additionally, a more detailed annotation guideline has been prepared, offering further explanations, examples, and counterexamples for each label. The guideline was made available to annotators to facilitate their understanding and ensure consistent application of the labeling criteria throughout the annotation process.

3.1.1 Patent title

The title of the patent application.
Example: VEHICLE SPEAKER DISPOSITION STRUCTURE

3.1.2 Section title

The title of each main section of a patent description.
Example: BACKGROUND ART

3.1.3 Section subtitle

The title of sub-sections inside the main sections of a patent description, if any.
Example: Stability studies

3.1.4 Technical field

Sentences determining the technical scope of the invention. These sentences specify to which field the invention relates and are usually carried out in one single paragraph.
Example: This application relates to the field of electronic materials and component technologies, and in particular, to an embedded substrate and a method for manufacturing an embedded substrate.

3.1.5 Reference

Sentences introducing the state-of-the-art or presenting the context to reach the invention, including related patents or publications, previous techniques, or general knowledge.
Example: In Patent Document 1, the acoustic transducer is disposed in the fender located near a front corner of a vehicle cabin, and sound is reproduced from the vicinity of the front corner toward the vehicle cabin.

3.1.6 Reference problem

Sentences stating the disadvantages of prior arts or indicating the technical problem that the invention is designed to solve.
Example: It can be learned that according to the existing embedded component packaging process, laser generated when the drill holes 104 are drilled damage the chip.

3.1.7 Reference advantage

Sentences explaining the advantage or quality of the prior arts or known technologies. In the example below, the first sentence provides context and the second sentence should be tagged as Reference advantage.
Example: In Patent Document 1, the acoustic transducer is disposed in the fender located near a front corner of a vehicle cabin, and sound is reproduced from the vicinity of the front corner toward the vehicle cabin. **By employing such a structure, an improvement in the reproduction efficiency, of high-quality sound including a low range, with a wide range of directivity in a plan view, is expected**.

3.1.8 Embodiment

Sentences describing physical instances or variations of the invention, explaining necessary ways to achieve the desired outcome. These sentences serve to demonstrate the flexibility and applicability of the invention in various contexts. (We keep the original reference numerals such as "104" in the text.)
Example: That is, a metal boss maybe disposed on each pad, and then embedded packaging (including drilling, conductive material filling, conductive layer disposing, and the like) is performed on the chip.

3.1.9 Invention advantage

Sentences providing the advantage, quality, or improvement brought about by the invention.
Example: The technique disclosed herein achieves both an improvement in the reproduction efficiency of the speaker and a reduction in the noise caused in the vehicle body by the sound generated by the speaker.

3.1.10 Invention problem

Sentences highlighting drawbacks or problems that the invention may cause.
Example: In short, since the speaker box 10 needs to have sealing properties in view of improving the reproduction efficiency of the sound including the low range, the drainage performance tends to deteriorate.

3.1.11 Figure description

Some patent applications contain figures which give a visual representation of the invention in the form of drawings, diagrams, or flowcharts. The Figure description tag is assigned to sentences that provide detailed explanations of figures, which usually contain reference numerals of the invention's components. These sentences should allow readers to navigate and understand the various depicted elements.
Example: As seen from Figure 3, for example, in its closed position, the door 20 is received within the recess 10 of the housing base 12 and its lower

face 24 lies generally flush with the lower surface 26 of the lip 14.

3.1.12 Definition

The explanation or clarification of technical terms, which could be a specifically formulated term. Context-specific explanations of which are given within the scope of the patent.
Example: As used herein, the term "cofactor" refers to a non-protein compound that operates in combination with a ketoredutase enzyme.

3.1.13 Rephrased claim

Sentences repeating portions of claims with non-substantive modifications, i.e., without incorporating additional content words that may alter the scope of the claims.
Example: A harness system for a power drive unit is disclosed. In various embodiments, the harness system includes an electrical cable having a first end and a second end, a plurality of cover members positioned along a length of the cable and a spring member positioned adjacent the plurality of cover members along the length of the cable.
original claim: A harness system for a power drive unit, comprising: an electrical cable (580) having a first end and a second end; a plurality of cover members (581) positioned along a length of the electrical cable; and a spring member (582) positioned adjacent the plurality of cover members along the length of the electrical cable.

3.1.14 Juridical template

Standardized phrases or sentences, which can be used regardless of the patent content. They serve specific purposes, such as facilitating transitions between sections of the description, and extend or narrow down the scope of the claims.
Example: The foregoing implementations of the present invention do not constitute a limitation on the protection scope of the present invention.

3.1.15 Technical template

Sentences giving the comprehensive usage of a technical term by providing its closely related synonyms or hyponyms.
Example: The first and the second plastic material may also be selected from a third group comprising a High Density Polyethylene, Low Density Polyethylene, Polyethylene, Terephthalate, Polyvinyl Chloride, Polycarbonate, Polypropylene, Polystyrene, Fluorine Treated, Post Consumer Resin, K-Resin, Bio-plastic, or combinations thereof.

3.1.16 Other

Sentences belonging to none of the previous categories or contains ambiguity. The following example demonstrates a typical OCR problem that has grouped all the elements of a table of contents together. Given that each title appeared in the corresponding subsection, this sequence is considered as *Other* to avoid introducing noise into the data.
Example: I. OverviewII. Description of StepsA. Tissue PreparationB. Distribution of DNA moleculesC. Detection and Quantification1. Digital PCR Methods2. Bead emulsion PCR3. Microfluidic Dilution with PCR4. Single molecule detection and/or sequencingD.

4 Corpus annotation

4.1 Corpus preparation

To build our annotation corpus, we use patent applications published by the European Patent Office (EPO). These patent applications are classified using the Cooperative Patent Classification (CPC) system, which comprises eight domains (Table 1). We randomly selected 2 applications per domain and divided them into two datasets, each containing one document per CPC class. In cases where an application is classified under multiple domains, we only considered the first one (the primary CPC label). The aim of this dataset separation is to verify whether the inter-annotator agreement remains consistent across the two datasets.

We extracted the description section from each patent application, each description is then segmented into sentences before being annotated. We use scispacy[1] combining with special rules to perform sentence splitting. For example, patent claims, which are usually extremely long sentences separated by semicolons, could be copied into description. In order to balance the size of each sentence, semicolons are also considered as ending punctuation. We chose to stay on the sentence level because delving into a finer-grained level would require not only knowledge in linguistics but also expertise in various technical domains. For instance, the following sentence, which we simply classify as being of type *Figure description*, could be broken down into sub-sequences that detail the interaction

[1] https://allenai.github.io/scispacy/

between elements described in the figure: *CPU 16 controls first conveyance section 21 to move conveyance pallet 40 to the loading position, and controls multi-joint robot 24 so that robot-side attachment section 27 grips pallet-side attachment section 42 below conveyance pallet 40 (refer to fig. 8).*

- Controller action: *CPU 16 controls first conveyance section 21*; *and controls multi-joint robot 24*
- Result of the action: *to move conveyance pallet 40 to the loading position*; *so that robot-side attachment section 27 grips pallet-side attachment section 42*
- Location of the action: *below conveyance pallet 40*
- Figure reference: *refer to fig. 8*

As an exploratory study and considering the number of defined labels, we decided to remain at the coarse-grained level. Table 1 shows the number of sentences in each document for both datasets of the corpus.

CPC	dataset 1	dataset 2
Human necessities (A)	393	228
Performing operations; transporting (B)	217	101
Chemistry; metallurgy (C)	349	681
Textiles; paper (D)	307	106
Fixed constructions (E)	364	102
Mechanical engineering; lighting; heating; weapons; blasting engines or pumps (F)	109	245
Physics (G)	224	284
Electricity (H)	193	221
Total (nb tokens)	62203	56886
Average tokens per sentence	*28.9*	*28.9*
Total (nb sentences)	**2156**	**1968**

Table 1: Number of sentences for each domain in the corpus (and the total number of tokens as well as the average tokens per sentence for information).

4.2 Annotation process

In order to measure the consistency across the annotation process, the annotation is conducted in two sessions. For each session, a pair of annotators (with a computational linguists background) independently annotate the same documents. Only one tag is allowed for each sentence. Discussion before the annotation is permitted, in order to allow both annotators to become familiar with the general structure of patent description. During the annotation, discussions are not allowed, instead, annotators have access to the context and any other information necessary for understanding the sentence to be annotated. After the first session, a collective review of the annotation guideline is conducted in order to complete the guideline with newly encountered examples.

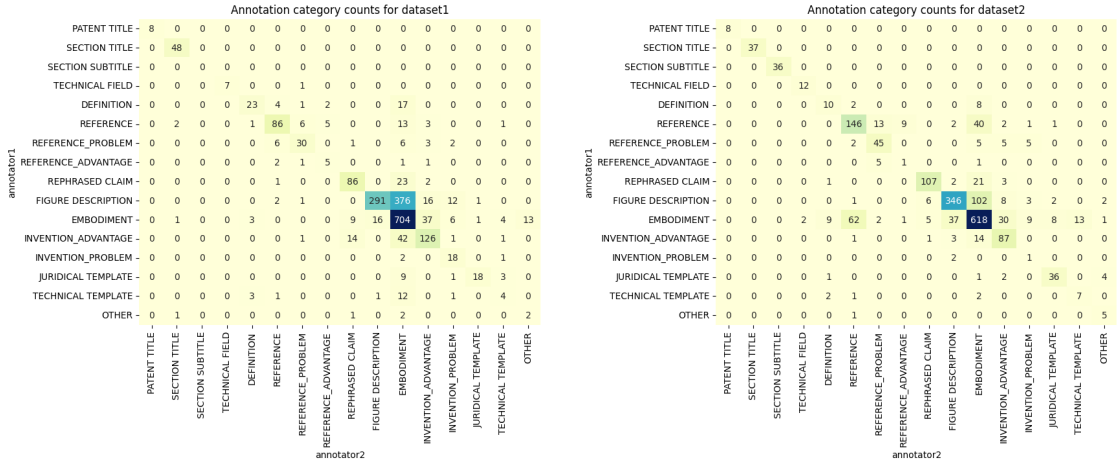
4.3 Annotation agreement

We employed the pairwise Cohen's kappa to measure inter-annotator agreement. Table 2 shows the scores for each class within each corpus. As explained in Section 3, the label *Section subtitle* was added after the annotation of the first dataset.

Labels	dataset 1	dataseet 2
Patent title	1.00	1.00
Section title	0.96	1.00
Section subtitle		1.00
Technical field	0.93	0.92
Definition	0.59	0.46
Reference	0.77	0.64
Reference_problem	0.67	0.70
Reference_advantage	0.45	0.11
Rephrased claim	0.76	0.84
Figure description	0.47	0.75
Embodiment	0.47	0.61
Invention_advantage	0.64	0.70
Invention_problem	0.58	0.09
Juridical template	0.70	0.79
Technical template	0.22	0.43
Other	0.19	0.55
kappa	**0.56**	**0.69**

Table 2: IAA for each label in each dataset of the corpus. The Cohen's kappa score for the entire dataset is used instead of the mean of scores for each category due to the imbalanced distribution of labels.

This modification has contributed to the perfect agreement concerning the labels associated with titles. It is worth noting that the agreement is relatively low for some labels, which is due to their imbalanced distribution in the corpus (as shown in Figure 1). The matrices in Figure 1 present the



(a) dataset 1 (b) dataset 2

Figure 1: The annotation by class across the two datasets. The matrices show, for each annotator, occurrence of labels assigned by each annotator for each dataset. We can observe that the two datasets are mainly composed of *Rephrased claim*, *Figure description*, *Embodiment*, and *Invention advantage*. In the dataset 2, *Reference* also represent a significant portion.

number of labels assigned to each sentence by each annotator. It can be observed that the majority of annotations fall under the categories of Figure description, Embodiment and Rephrased claim, followed by Reference and Invention advantage. This is consistent with the objective of patent description drafting, which aims to explicitly explain the way of carrying out the invention and its novelty compared to prior arts.

Overall, we observe an improvement in the agreement between the two annotation sessions, particularly for *Figure description* and *Embodiment*. The label *Other* denotes sentences that appear ambiguous or do not belong to any class. Few sentences received this label from both annotators, which shows that our label set is comprehensive enough to cover the entirety of sentences in a patent description.

As we can see from IAA score and Figure 1, apart from labels concerning titles, the categories that receive higher agreement are those less dependent on the annotator's interpretation of sentence meanings, such as *Technical field*, *Rephrased claim*, and *Juridical template*.

Compared to the work of (Nakamitsu et al., 2022), we have expanded the label set with the intention to encompass the entirety of patent description content, rather than merely focusing on specific parts thereof. We attempted to create an exemplary dataset manually, with the objective of identifying relevant labels and establishing a reliable sample for future data augmentation. On the whole, our annotation achieved an IAA score of 0.69 following a revision on the first dataset. This result aligns with the performance of similar annotation work (Fisas et al., 2015), who obtained an IAA score of 0.6567 across eight categories.

However, although we improved the agreement score for some categories, it is worth noting that the level of agreement remains relatively low for some categories, despite post-annotation discussions following the first session. The following section gives examples of pairs of labels that are frequently confused by annotators.

4.4 Disagreement analysis

Based on the annotation results, we noticed that certain pairs of labels are often confused. We attempted to analyze the reason for this confusion.

4.4.1 Reference VS Embodiment

The challenge associated with this pair of labels lies in distinguishing whether the subject of the sentence concerns the prior art or the applicant's invention. The reason is that, in the section describing embodiments of the invention, the description of the invention can be mixed with the explanation of prior art. This is particularly the case for patents in the *Chemistry; metallurgy* domain. In these patents, the disclosure of detailed experimentation is required, which often leads to numerous references when existing components or methods are needed for the experiments. Consequently, the

description of the invention example becomes intertwined with that of the prior art, complicating the annotators' comprehension, especially when they lack domain-specific expertise.

Example: Fluorescent nucleotide incorporation by DNA polymerase. As described in the above-referenced PNAS publication by Braslavsky et al., DNA polymerase may be employed to image sequence information in a single DNA template as its complementary strand is synthesized. The nucleotides are inserted sequentially;

In this example, it is challenging to determine whether the sentence underlined is part of the publication cited in the previous sentence or merely a step in the *Fluorescent nucleotide incorporation by DNA polymerase* experiment. We have chosen *Embodiment* as the label for this sentence because the following context explains the experiments related to the invention itself and not the reference.

4.4.2 Reference advantage VS Reference problem

Distinguishing between advantages and problems can be challenging, especially when purely critical or, conversely, commendatory terms are missing.

Example: Furthermore, in regular operation, an auxiliary circuit may be energized and connected to a junction by way of a second current interrupting element. Electrical power can thus be provided from DFIG to auxiliary components, with the electrical power from main power transformer being converted to the appropriate voltage by auxiliary transformer. However, during maintenance operations, the DFIG may be shut down, and the main power circuit may be isolated from the power grid.

In this example, it's difficult to tell if *DFIG may be shut down* and *main power circuit may be isolated* are positive characteristics even though the two preceding sentences have provided context. With the help of the following context, we understood that it indeed represents an advantage, especially it was mentioned that this can help to *reducing the risk of electrocution during maintenance operations.* We thereby decided to annotate it as *Reference advantage*.

4.4.3 Reference problem VS Invention advantage

Sometimes, the information is presented as a dual statement which requires annotators to interpret the context and infer the intended meaning.

Example: This entails the need to exert a high torque by the motor to carry out the movement quickly.

In this example, it can be inferred that the sentence implies a drawback of the current technique. However, in a patent description, such a sentence exists only to indicates that the mentioned problem will be rectified through the invention, thereby expressing an advantage of the latter. To solve the ambiguity, we added a rule to our annotation guideline, explicitly stating that for such dual statements, the sentence will be annotated as *Invention advantage* because the presented problem would be solved by the invention.

4.4.4 Embodiment VS Figure description

Despite the introduction of additional specifications after the first annotation session regarding the distinction between *Embodiment* and *Figure description*, with a particular emphasis on the functional aspect of the former and the visual aspect of the latter, the differentiation remains challenging. This is because the description of an embodiment often refers to components drawn in the figures.

Example: In atmospheric pressure plasma-generating device 10, processing gas composed only of an inert gas is supplied from first connecting passage 130 to reaction chamber 100 through the inside of holders 72 and 74 of holding member 20.

In this example, the technical terms followed by numbers indicate that these are important components of the invention and that they are illustrated in the drawings. However, the sentence only describes how the processing gas is supplied, which is not depicted in the drawings. Considering the process is not shown in the drawings, we decided to label it as *Embodiment* and we clarified that it is possible to refer to drawings attached to the patent applications in case of indecision between *Embodiment* and *Figure description*.

4.4.5 Embodiment VS Invention advantage

Using comparatives when describing an invention may not always clearly indicate an improvement of the invention. The confusion often caused by insufficient technical knowledge in the respective domain.

Example: It is therefore known that the particle size distribution computed using the profile data about the coke 30 shows larger particle size distribution than the actual particle size distribution.

In this example, it's difficult to decide if *larger particle size distribution* is an improvement

achieved by *using the profile data about the coke 30*. Thus, in cases where we cannot be certain that the presented feature is an advantage, we annotate it as *Embodiment* in order to avoid introducing errors.

The analysis of disagreement sheds light on the challenges involved in annotating the discursive roles of sentences in patent descriptions, which are not only related to language complexity but also to individual manner of expression.

5 Conclusion

In this paper, we have proposed an annotation scheme adapted to the specific writing style of patent applications. As an exploratory work, we defined a set of 16 labels to categorize each sentence in a patent description according to their discursive roles. The initial results show that such work is feasible, since strong agreement is achieved for most categories. However, challenges remain. Considering the aforementioned difficulties, we propose the following improvement to the future work: allow multi-labeling for ambiguous sentences or consider implementing a multi-layer annotation scheme. In the first level, include classes corresponding only to the five common sections of a patent description, followed by additional specific categories in subsequent layers as necessary.

In conclusion, our annotation work is an ongoing process. We plan to expand our dataset once the relevant labels have been established, and to employ active learning methods to streamline the annotation process. We believe that a such linguistic resource in the patent domain could contribute to enhancing the accuracy of tasks such as patent classification, patent novelty detection, patent information retrieval, and, most central to Qatent, computer-assisted patent drafting.

Limitations

Our sample dataset contains only 16 randomly selected documents, which might not be sufficient to contribute to classification model training. Additionally, our work could benefit from having a lead annotator to supervise the annotation process. This would help reduce the time spent on correcting annotation errors and ensure adherence to the annotation guideline.

Ethics Statement

All data used in this study were collected from publicly available sources, with no infringement of intellectual property rights or privacy.

Acknowledgments

We would like to express our profound thanks to You Zuo for taking the time to read our paper, offering thoughtful comments, and providing enriching ideas. Finally, We are grateful to the anonymous reviewers and for their detailed and helpful feedback.

References

Pradeep Dasigi, Gully Burns, and Anita Waard. 2017. Experiment segmentation in scientific discourse as clause-level structured prediction using recurrent neural networks.

Gabriela Ferraro, Hanna Suominen, and Jaume Nualart. 2014. Segmentation of patent claims for improving their readability. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 66–73, Gothenburg, Sweden. Association for Computational Linguistics.

Beatriz Fisas, Horacio Saggion, and Francesco Ronzano. 2015. On the discoursive structure of computer graphics research papers. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 42–51, Denver, Colorado, USA. Association for Computational Linguistics.

Jieh-Sheng Lee and Jieh Hsiang. 2020. Patent classification by fine-tuning bert language model. *World Patent Information*, 61:101965.

Xiangci Li, Gully Burns, and Nanyun Peng. 2021. Scientific discourse tagging for evidence extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2550–2562, Online. Association for Computational Linguistics.

Jun Nakamitsu, Satoshi Fukuda, and Hidetsugu Nanba. 2022. Analyzing the structure of u.s. patents using patent families. In *2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 150–153.

Masayuki Okamoto, Zifei Shan, and Ryohei Orihara. 2017. Applying information extraction for patent structure analysis. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 989–992, New York, NY, USA. Association for Computing Machinery.

L. Siddharth, Lucienne T. M. Blessing, Kristin L. Wood, and Jianxi Luo. 2021. Engineering Knowledge Graph From Patent Database. *Journal of Computing and Information Science in Engineering*, 22(2). 021008.

WIPO. 2022. Wipo patent drafting manual, second edition. page 97.

H. Zuo, Y. Yin, and P. Childs. 2022. Patent-kg: Patent knowledge graph extraction for engineering design. *Proceedings of the Design Society*, 2:821–830.