Automatic Identification of Patent Claim Types: Enhancing Efficiency in Patent Analysis

Rima Dessí[1], Hidir Aras[1], Mark Prince[2] and René Hackl-Sommer[1]

[1]FIZ Karlsruhe - Leibniz Institute for Information Infrastructure, Germany
[2]CAS - Chemical Abstracts Service, Columbus, Ohio, USA

Abstract

Patents are an important source of technological innovation. Due to the large number of patents published each year, it has become increasingly difficult to find precise information on specific patent inventions, which requires not only professional search systems but also many years of patent expertise. Patent claims are the backbone of inventions and define the scope of legal protection. They can be classified in different types in terms of what they claim, e.g. for a physical entity we can speak of "product claims", while for an activity we can refer to as a "process claim". Manual identification of these claim types for a large set of documents is labor-intensive and time-consuming. To address this challenge, we developed a Patent Claim Type Recognition (PCTR) model based on Deep Learning (DL), which is able to automatically identify pre-defined types of patent claims. Further, we also built a rule-based heuristic approach, to generate training data to be used by the PCTR model. The proposed model was evaluated by using a dataset labeled by Subject Matter Experts (SMEs). Our experimental results demonstrate that the PCTR model accurately identifies the type of given patent claims, offering a promising approach to streamline patent analysis and evaluation processes.

1. Introduction

Patents enable inventors to disclose their inventions and protect them legally by preventing others from using, selling, and producing the invention without permission [1]. Therefore, patents encourage further inventions by granting exclusive rights to inventors and thus foster more research and development. However, the ever-increasing number of available patents and their complex characteristics in nature poses several challenges to scientists, lawyers and information professionals. These documents are diverse, encompassing text, formulas, drawings, tables, and more, while also being lengthy and filled with domain-specific vocabulary that is tailored to the target field. The so-called full text of a patent document often consists of sections, namely, title, abstract, claims, and description.

The claims are a crucial component of the patent documents, they define the legal scope of protection of an invention. Essentially, they specify the subject matter that is sought to be protected and refer to the core inventive information of a patent. These claims serve as the foundation of what aspects of an invention should be protected from infringement. Therefore, an accurate analysis and understanding of them are vital for inventors, examiners, and scientists. There can be independent and dependent claims, together building a hierarchy[1]. Typically, independent claims contain core inventive information whereas dependent claims specify improvements or variations. Such variations can be rather miniscule or they can be quite substantive. However, the legal jargon in patent claims can make it difficult to understand what exactly a text is about. Further, the claims often relate to a particular subject matter, such as apparatus, composition, process, or a combination thereof. Therefore, it is crucial to precisely identify the type of a claim for accurate patent analysis. However, manual identification of claim types in larger result sets is expensive and time-consuming.

Several studies have been proposed to efficiently and effectively analyze and understand patents as well as their claims. Most of these methods employ Machine Learning (ML) and Natural Language Processing (NLP) techniques to automate patent classification, claim classification, and claim type identification. However, often they require manually labeled large amounts of training data. Further, approaches focus on claims designed to measure different aspects of patents such as comparison of patent claims and economic growth and social welfare [2] or to measure technological patent scope with semantic analysis of patent claims [3].

In this paper, we propose a Patent Claim Type Recognition (PCTR) model based on Deep Learning (DL) techniques. As mentioned before, patent claims refer to a particular subject matter, i.e., types such as apparatus, composition, system, etc. The main goal of PCTR is to automatically identify this type information for each given

5th Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech) 2024
rima.dessi@fiz-karlsruhe.de (R. Dessí); hidir.aras@fiz-karlsruhe.de (H. Aras); mprince@cas.org (M. Prince); rhacklsommer@gmail.com (R. Hackl-Sommer)
0000-0001-8332-7241 (R. Dessí); 0000-0002-3117-4885 (H. Aras)

[1]https://www.wipo.int/edocs/mdocs/aspac/en/wipo_ip_phl_16/wipo_ip_phl_16_t5.pdf

patent claim. Additionally, we design a rule-based heuristic model to generate training data for the PCTR model. The training data consists of claims paired with their respective types, determined through the heuristic model. Subsequently, this curated data is used to train the PCTR model, finally, the trained model is able to assign a claim type to a given patent claim.

Overall the main contributions of the paper are as follows:

- Introducing a rule-based heuristic model that assigns types to claims, facilitating the generation of training data.
- A transformer-based deep neural network architecture designed to automatically identify patent claim types.
- A comprehensive evaluation of the PCTR model using data labeled by three different Subject Matter Experts (SMEs).

2. Related Work

Due to the significant importance of patent documents for individuals, enterprises, and industries, there has been a considerable amount of study and research dedicated to this domain. These studies cover a wide range of topics such as patent classification [4], patent landscaping [5], prior art searches [6], and more. Further, since claims are the core part of patent documents that define the legal boundaries to protect inventions, there has been a concerted effort to utilize them effectively to propose scientific solutions.

As mentioned in the Introduction section patent claims can have different types, [7] analyzes the occurrence of *process* claims in large US patent corpus and reports that substantial increase over the last century in such type of claims. Further, the authors also developed a patent claim classification tool[2] that recognizes three types of claims namely, process claims, product claims, and product-by-process claims. [8] investigates the relation between the patent examination process and the patent's scope. The proposed method relies on the claim length and count. Another interesting study performed by [9], in which authors first collect patents from different countries to assess the country's technological capability by comparing the number of patents and claims. It turns out that patent claims are much more reliable than the number of patents to reflect the country's technological advancement.

In contrast to these approaches, our proposed approach differs in two main aspects. First, the focus of our work is to build a comprehensive model to identify types of pre-defined patent claims. Second, although our method is based on a deep neural network that includes a transformer layer, it does not require any manually labeled data, instead, we define a heuristic model to label large amounts of data efficiently and effectively without requiring any manual effort. Consequently, this dataset is then used to train the proposed PCTR model.

[2]https://zenodo.org/records/6395308

3. Patent Claim Type Recognition

In this section, we give a definition for the regarded problem and describe the predefined claim types for the model prediction.

Problem Definition: Given a claim text and a predefined type list, the task of the PCTR model is to assign the most relevant type from the predefined type list.

Predefined Types: The specific claim types listed below, which are defined by name and example in the WIPO[3] publicly available documentation, focusing on the subset of claim types that are directed at the nature of the invention.

- **Method:** recites a sequence of steps that complete a task or accomplish a result
- **Use:** depicts intended/inventive application of novelty
- **Composition:** invention pertains to the chemical nature of materials/components used.
- **Process:** claims define a process of manufacture, it should be noted that WIPO documentation labels it as "product-by-process".
- **Apparatus:** protects an apparatus or device
- **System:** an assemblage or combination of things or parts forming a unitary whole

To automatically classify a given claim into the above-described types we used a (rule-based) heuristic method (see section 4.1) to generate training data. This data is then used to train the Patent Claim Type Recognition (PCTR) model which is based on a deep neural network for automatic claim type identification. Figure 1 illustrates the architectural design of the PCTR model.

PCTR is a transformer-based multi-class classification model that is capable of assigning the most relevant type to a given patent accurately. Figure 1 illustrates the claim type recognition model, i.e., the deep neural network model that has been designed for this study. It consists of a transformer block which is integrated as a layer, followed by a pooling layer, a dense layer, and a final softmax layer. The input to the model comprises a claim paired with the document sections such as the title and abstract. Then the output is the type of the given claim, represented as $P(y = t|X)$, where y denotes the patent

[3]https://www.wipo.int/edocs/mdocs/aspac/en/wipo_ip_phl_16/wipo_ip_phl_16_t5.pdf

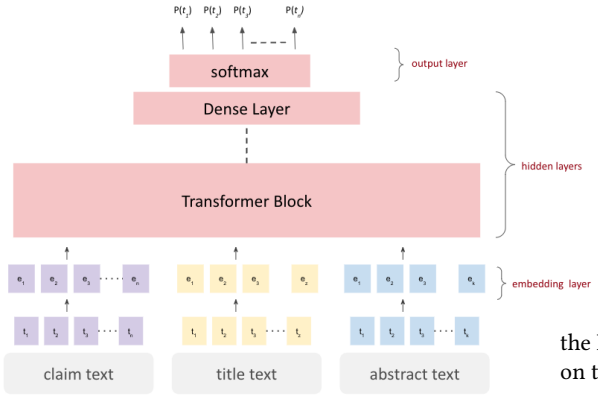

Figure 1: The general architectural overview of the PCTR model.

Claim Type	#Sample
System	322,848
Process	84,847
Method	988,687
Composition	141,273
Apparatus	540,025
Use	14,705

Table 1: Statistics of the Labeled Training Data

claim type (e.g., apparatus, composition, system, etc.). The model aims to classify the claims based on the provided input text. Figure 1 illustrates an example input of a patent section and its associated claim. Initially, the text is tokenized, and the token embeddings serve as input to the transformer block, with these embeddings being randomly initialized.

Classifying patent claims according to given types (as defined above) is a challenging task. It should be noted that in this study, we distinguish claim types based on their level of complexity as *simple claim* types and *complex claim* types. Simple claim types, as the name suggests, are straightforward and include claim-type information within the claim text. On the other hand, complex claim types either do not include explicit claim-type information in the claim text or refer to more than one type, making it challenging to identify them. The focus of the developed Patent Claim Type Recognition (PCTR) approach is to automatically identify the type of complex claims by designing and developing the deep neural network.

3.1. Feature Selection

To train the PCTR model, we primarily utilized textual features, including the claim, title, and abstract. Other potential features, such as CPC/IPC[4] information and additional contextual or structural elements from the patent document, remain for our future analysis.

To train the PCTR model with different feature combinations we designed the following versions of it: PCTR_V1 utilizes claim, PCTR_V2 utilizes claim + title, PCTR_V3 utilizes claim + title + abstract as input to perform the claim type prediction. Essentially, the version of the PCTR model (i.e., V1, V2, or V3) is determined based on the features that have been exploited.

[4]https://www.epo.org/en/searching-for-patents/helpful-resources/first-time-here/classification

4. Experimental Results

This section gives an overview of the generation of the training data, test data annotated by subject matter experts (SMEs), and the experimental outcomes obtained from the PCTR models.

4.1. Training Data Generation

To train the PCTR models which are based on the DL architecture, it is crucial to have a sufficient amount of representative training data. Manually, generating millions of labeled data is expensive and time-consuming. For the training data generation only simple claims (cf. Section3), which contain the type information within the text are considered. We designed a rule-based heuristic model that is able to label given simple claim texts based on a defined regular expression (regex) rule. On the other hand, the test set contains only complex claims (cf. Section4.2), which do not include the claim type within the text.

The designed regex baseline relies on start and end markers. In between such markers are the targets. Following is an example of the beginning of a patent claim:

Example 1. *"68. A method according to claim 67," with "68. A" being a start marker, "according to" an end marker, and "method" the target.*

End markers are different depending on whether they are dependent or independent claims. Therefore, such information has to be provided, e.g. by using different tools designed for this purpose. For that, the internally developed tool has been employed. Three types of targets are identified. High-probability targets where start and end markers only capture a single word and that word matches one of the known claim types (apparatus, compound, composition, device, method, process, system, use). Medium-probability targets where several

Claim Type	#Sample
System	98
Process	37
Method	83
Composition	152
Apparatus	89
Use	31

Table 2: Statistics of the Labeled Test Data

Model Type	Features	Accuracy
PCTR_V1	claim	0.561
PCTR_V2	claim + title	0.576
PCTR_V3	claim + title + abstract	0.602

Table 3: Performance of PCTR Models

Model Type	Features	Accuracy incl. feedback
PCTR_V1	claim	0.808
PCTR_V2	claim + title	0.816
PCTR_V3	claim + title + abstract	0.824

Table 4: Performance of PCTR Models after including SMEs feedback

words a captured, but one of them is a known claim type. And lastly, low-probability targets, where more than one known claim type is captured.

After applying the heuristic model to the claim texts of internal sources which encompasses patents from two different sources, namely, the World Intellectual Property Organization (WIPO), and the US Patent Office. 2 million patent samples were selected employing random sampling and subjected to preprocessing. In other words, a subset of data points has been selected as a training set from a larger dataset. The preprocessing involved removing duplicates as well as invalid claim types, abstracts, and titles. Finally, we had 2,092,385 samples with high probability targets, i.e., claim types. All the claims with low-probability or medium-probability targets are filtered. The statistics of the training data are shown in Table 1.

It should be noted that to avoid biases towards the rule-based method for labeling training samples and allow for the generalization of the trained models, we removed the claim type information from each patent claim before feeding them into the PCTR models.

4.2. Test Data Generation

To accurately assess the effectiveness of the PCTR model, it is essential to have representative test data that the model is expected to encounter in real-world scenarios. To this end, the test data was generated and labeled by overall 4 SMEs following an iterative process in order to achieve the required level of agreement. Initially, 300 samples were randomly selected from the two internal data sources to be labeled by the SMEs. Additionally, 200 samples were also selected from an external database to balance the test set, ensuring an equal number of expert-labeled samples for each type. The statistics of the test set are presented in Table2. For determining the final claim type, we adopted a majority vote approach.

4.3. Evaluation of PCTR Models

We have trained 3 different PCTR models namely: PCTR_V1, PCTR_V2, PCTR_V3. Table 3 presents the accuracy results of the models, calculated as the ratio of correctly classified data to the total test data. Upon analyzing these results and consulting with SMEs for improvement suggestions, we received valuable feedback.

The analysis by SMEs led to a revised interpretation of the PCTR model's performance. Some samples initially deemed as falsely identified types were actually correctly identified, according to the feedback of the SMEs. The essential feedback that we applied to our evaluation process to improve the performance of the PCTR models is as follows:

- "Product-by-process" claims should be regarded as "process" claims.
- "Use" claims can be seen as a sub-category of "method" claims.
- Claims can have multiple (2) types.

Table 4 presents the improved accuracy of the PCTR models after considering this feedback and re-computing the accuracy.

Another aspect has been considered to improve the models' performance. The generated dataset through random sampling is quite unbalanced and generally, it is a good practice to have a balanced data set for any machine learning model. In our efforts to achieve a balanced dataset, we had to reduce the size of the training data by downsampling the dataset as some claim types were underrepresented. However, this reduction in the training data resulted in a drop in accuracy, as there were fewer samples for each claim type hence a smaller dataset. Therefore, we decided to use the originally randomly sampled training set to train the models. It should be noted that ensuring a balanced training dataset while randomly sampling is a time-consuming process that requires expert assistance to ensure an equal number of training samples for each class or claim type.

5. Conclusion

In this paper, we presented an approach for the automatic identification of predefined claim types based on a Deep Learning model. Looking at the performance of the PCTR model's performance focusing on both dataset characteristics and feature selection aspects, we can report the following results of a deeper analysis.

Data from different Patent Authorities: As claim type definitions have been discovered to vary by jurisdiction, it is needed to customize the developed models per patent office (or collections of patent offices) accordingly. As stated earlier, the claim type definitions are based on WIPO documentation.

The importance of feature selection: The experiments indicate that including more features, specifically the title, and abstract, can significantly improve the accuracy of the PCTR models for claim type identification. In general, providing more context that can be used to describe and distinguish each claim type from each other is helpful. Further, our preliminary experiments with a very small set of datasets suggest that including CPC information, improved the accuracy. Nevertheless, including CPC as a feature requires a systematic evaluation and data sampling considering the various domains and extraction of balanced data of sufficient size for each claim type. We leave this as our future work.

The role of SMEs: The results also demonstrate that incorporating the feedback and insights from SMEs can greatly enhance the accuracy of the PCTR models. This highlights the importance of involving subject matter experts in the development process. Herewith, starting from investigated examples in our data analysis it was, for example, confirmed that a claim can be assigned several types. Besides that, it turned out that it is viable to consider several probabilities for the correct target label (type), as for example in the case of product-by-process claims which were predicted widely as process. With this valuable feedback, we plan to further refine the PCTR implementation to allow for assigning multiple type information in future iterations.

Finally, the developed PCTR model can be applied in various real-world scenarios: (1) as a stand-alone model targeting only complex type claims, or (2) as part of a hybrid system combining a heuristic model (cf. Section 4.1) with the PCTR model.

References

[1] R. Dessi, H. Aras, M. Alam, Exploring the impact of negative sampling on patent citation recommendation, PatentSemTech (2023).

[2] S. Niwa, Patent claims and economic growth, Economic Modelling 54 (2016) 377–381.

[3] S. Wittfoth, Measuring technological patent scope by semantic analysis of patent claims–an indicator for valuating patents, World Patent Information 58 (2019) 101906.

[4] C. J. Fall, A. Törcsvári, K. Benzineb, G. Karetka, Automated categorization in the international patent classification, in: Acm Sigir Forum, volume 37, ACM New York, NY, USA, 2003, pp. 10–25.

[5] S. Choi, H. Lee, E. Park, S. Choi, Deep learning for patent landscaping using transformer and graph embedding, Technological Forecasting and Social Change 175 (2022) 121413.

[6] S. Bashir, A. Rauber, Improving retrievability of patents in prior-art search, in: European Conference on Information Retrieval, Springer, 2010, pp. 457–470.

[7] B. Ganglmair, W. K. Robinson, M. Seeligson, The rise of process claims: Evidence from a century of us patents, ZEW-Centre for European Economic Research Discussion Paper (2022).

[8] A. C. Marco, J. D. Sarnoff, A. Charles, Patent claims and patent scope, Research Policy 48 (2019) 103790.

[9] X. Tong, J. D. Frame, Measuring national technological performance with patent claims data, Research policy 23 (1994) 133–141.