Automating Claim Construction in Patent Applications: The CMUmine Dataset

Ozan K. Tonguz Yiwei Qin Yimeng Gu Hyun (Hannah) Moon
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213-3890, USA
{tonguz, yiweiq, yimengg}@andrew.cmu.edu, hyunm@alumni.cmu.edu

Abstract

Intellectual Property (IP) in the form of issued patents is a critical and very desirable element of innovation in high-tech. In this position paper, we explore the possibility of automating the legal task of Claim Construction in patent applications via Natural Language Processing (NLP) and Machine Learning (ML). To this end, we first create a large dataset known as CMUmine™and then demonstrate that, using NLP and ML techniques the Claim Construction in patent applications, a crucial legal task currently performed by IP attorneys, can be automated. To the best of our knowledge, this is the first public patent application dataset. Our results look very promising in automating the patent application process.

1 Introduction

In the USA, European Union (EU), and Asia, most of the high-tech industries (semiconductors, wireless, Internet, Telecommunications, Robotics, Sensors, etc.) are characterized by the innovations they introduce and the inventions they make in a specific area of technology and these innovations are considered to be the intellectual property (IP) of these companies, a very important asset for any high-tech company. To protect their IP, high-tech companies file for patent applications to make it official that they own that specific idea and invention that could involve a new system or apparatus, new method, and new algorithms and/or software.

In general, filing patents that describe new inventions is a lengthy, cumbersome, and very costly process since most high-tech companies have to hire law firms, litigation attorneys who specialize in IP, and technical professionals called patent "agents", or "patent engineers" for writing and filing such patent applications. After an inventor prepares a well-written document describing his/her invention, she/he submits this document to an IP attorney who takes this document that is known as “Invention Disclosure” and prepares a patent application that can be submitted in the US to the United States Patent and Trademark Office (USPTO).

The main sections of a patent comprise the following sections:

1) Title and Inventors
2) Abstract
3) Introduction (background of the invention and “prior art”)
4) Invention Summary
5) Description of the invention including figures (also known as “specifications”)
6) Claims (independent claims and dependent claims)

A key observation we make is that, if the Invention Disclosure is well prepared by the inventor, then the main contribution of the IP attorney to the patent application is to formulate the claims of the specific invention. In legal terms, this task is known as "Claim Construction". An IP attorney will not typically change the other parts of an invention (e.g., the specifications section of the invention); instead, he will formulate the main claims of the invention in the form of:

(i) Independent Claims
(ii) Dependent Claims

and then append these claims to the end of the invention disclosure for the official submission of the patent application to the USPTO. Typically, patents in information technology have 3 Independent Claims and 5 or 6 dependent claims per independent claim that do depend on each of the 3 independent claims, thus resulting in a total of 20 or more claims in a patent application. In general, the independent claims concern the following aspects of an invention:

1) System or apparatus claim
2) Method claim
3) Software claim

Natural Legal Language Processing Workshop 2021, pages 205–209

2 Problem Statement

In this position paper, we explore the possibility of replacing the "human agents" with an automated solution (i.e., "the machine") in patent application process. In other words, we investigate whether the construction of the independent and dependent claims prepared by an IP attorney can be automated by using AI, Natural Language Processing (NLP), and Machine Learning (ML). The key observation behind this is the underlying pattern in preparing the claims in a patent application: the IP attorney gets a well written Invention Disclosure from an inventor and, based on that and a brief conversation with the inventor, prepares the claims of a patent application. Can this process be automated? Our results suggest that the use of NLP and ML can indeed automate "Claim Construction" tasks and, therefore, the patent application process. As a proof-of-concept, among all the claims constructed by IP attorneys, this paper focuses on generating the First Independent Claim. The approach we pursue in this paper is to formulate the problem as a text summarization problem [El-Kassas et al., 2021].

3 Related Work

Text summarization aims to briefly summarize the key information of any longer input text. Text summarization techniques using NLP has been successfully applied to various fields, including news [Grusky et al., 2018] [Fabbri et al., 2019], scientific papers [Lu, 2011] [Clement et al., 2019] and patents [Grusky et al., 2020]. Specifically, in [Grusky et al., 2020], the author models the text summarization task as follows: a granted patent's description is the input text, and its summary is regarded as the gold-standard summary. Our task is similar to other text summarization tasks, i.e., summarizing the First Independent Claim from a longer Invention Disclosure. Therefore, a viable approach that might work is to model our task as a text summarization task.

4 Data Collection and Dataset Construction

To meet our goal, we started by collecting a large dataset consisting of 300K patent applications, known as Carnegie Mellon University Machine Interpreted Natural-Language Engineering (**CMUmine™**). Using this very large data set, dubbed **CMUmine™**, of 317,356 previous US patent applications, we create a training set, a validation data set, and a test set that comprise 253,976, 31,736, and 31, 644 data points (i.e., previous patent applications), respectively. This is the first public and largest patent application dataset to the best of our knowledge. BigPatent [Grusky et al., 2020] is a well-known large text summarization dataset in patent domain, but it does not include patent claims and cannot be used for claim construction.

4.1 Description

Our dataset consists of issued patent applications collected from USPTO Bulk Data Storage System (BDSS) Version 1.1.0 [1]. Specifically, our raw data comes from the *Patent Application Full Text Data (No Images) (MAR 15, 2021 - PRESENT)* in the link above. We only used the patent applications issued in 2005 and 2006 to construct our dataset. Based on a detailed literature review on popular text summarization dataset's size, we decided to collect roughly 300,000 data points. Initially, we began to process data from the year 2002 and found that from the year 2002 to year 2004 datasets have non-standardized data structures. When we processed the year 2005 data set, it had a suitable structure for processing the data. It also met our expectations on the amount of data needed to train models after accumulating two years of data in a row. Therefore, we decided to use the year 2005 to 2006 data envisioning that the same arguments can be applied for data collected in recent years as well. Our data set can be found at this link[2].

4.2 Data Processing

The raw data on USPTO BDSS is in the format of XML, e.g., *ipa150903.xml*. Each xml file contains all the patent applications issued during that week of a certain year.

Considering the fact that patent applications are organized in different ways, to reduce the variation in size of summary and First Independent Claim and build a more representative dataset, we applied filter conditions to remove outliers. We only kept patent applications whose Invention Summary and First Independent Claim's length are within the percentile [10%,90%], that is [150,1500] words for

[1] https://bulkdata.uspto.gov/

[2] https://drive.google.com/drive/u/0/folders/1J4sAcM_21G39VuZT1jv6RqLTEM_UngWS

Invention Summary and [35, 300] words for First Independent Claim.

4.3 Dataset Structure

Our dataset consists of training set, validation set, and test set with a ratio of 8:1:1. Under train/validation/test folder, there are 6 subfolders: 1) abstract; 2) background; 3) summary; 4) detailed description; 5) first independent claim; and 6) claims. Under each sub-folder, every single file is named in the format of `{patent application No.}_{sub-folder name}`. All the files only contain text data. It was observed that not every US patent application contains a detailed description section. If a patent application does not have a detailed description section, we do not include it in the detailed description sub-folder. Around 2/3 (two thirds) of patent applications in our dataset have detailed description part.

4.4 Dataset Analysis

To have good insights into the features of our dataset, we use several automatic metrics to quantify its important features (e.g., average length, extractivity, compression, novel words).

Compression ratio [Grusky et al., 2020]: ratio between source document and output length. Compression ratio is measured by

$$CMP(S, O) = \frac{S}{O}, \quad (1)$$

where |S| and |O| denote the length of the source document and output sequence, respectively.

Coverage [Grusky et al., 2020]: measures the percentage of words in the output sequence that are part of an extractive fragment in the source document. Coverage is measured by

$$Coverage(S, O) = \frac{1}{O} \sum_{f \in F(S,O)} |f|, \quad (2)$$

where F (S, O) is the set of shared sequences of tokens in source S and output sequence O.

Density [Grusky et al., 2020]: measures the average length of the extractive fragment. Density is measured by

$$Density(S, O) = \frac{1}{|O|} \sum_{f \in F(S,O)} |f|^2. \quad (3)$$

Copy Length [Chen et al., 2020]: measures the average length of segments in output sequence copied from source document.

Novelty n-gram ratio [Narayan et al., 2018]: the proportion of segments in the output sequence that haven't appeared in source documents. The segments can be instantiated as n-grams.

As mentioned before, our dataset includes different parts of an invention disclosure document. Since the useful information to generate claims are scattered to different sections of this document, it is important to evaluate the features of different parts to decide which part is the best to use as the input sequence for our model to generate claims. For simplicity, we do not consider the combination of two parts and the detailed description part, whose length exceeds the capacity of both Recurrent Neural Networks (RNN) based model and Transformer [Vaswani et al., 2017] based model. We evaluate the dataset characteristics using Abstract, Introduction, or Invention Summary as source document and First Independent Claims as the output. The dataset evaluation result is shown in Table 1.

From Table 1, we observe that the summary part has the highest extractive rate, which is reflected in the highest density, coverage, copy length and lowest novelty words ratio. This means the summary part is the most informative part for generating claims, so we use the summary part as the model input in our work.

5 Approach

As described in Section 2, we formulate our problem as a text summarization task, so we use popular summarization systems: Pointer Generator (PG)[See et al., 2017] and PEGASUS [Zhang et al., 2019]. We used the training data set to train our model, the validation data set to fine-tune our model in terms of hyper parameters and then evaluated the performance of our model in generating the First Independent Claim (FIN) on the test set. The results obtained are compared with the "Gold Standard" that represent the versions of the same FIN constructed by humans (i.e., IP attorneys). The obtained machine-generated results are compared in terms of ROUGE-1, ROUGE-2, and ROUGE-L [Lin, 2004] scores based on the statistical distributions of the results obtained, i.e., the probability density functions (pdf's) of the results.

6 Results

6.1 Evaluation of the generated first claim

Table 2 reports F1 scores of ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL) for all models.

	Summary	Abstract	Background
Mean Source Length	615.5	123.2	726.6
Mean Output Length	123.8	123.8	123.8
Compression Ratio	4.97	0.995	5.87
Novelty1-gram/2-gram	0.20/0.38	0.36/0.58	0.53/0.85
Density	20.18	9.22	1.58
Coverage	0.86	0.73	0.65
Copy Length	9.01	4.41	1.55

Table 1: Intrinsic Characteristics of patent applications in the CMUmine dataset

	R1	R2	RL
Pointer Generator	65.51	52.95	59.25
PEGASUS	75.97	64.47	70.54

Table 2: Average F1 scores of the first independent claim generated with the test set for Rouge-1, Rouge-2, and Rouge-L.

It can be observed from Table 2 that the performance of self-attention approach is distinctively higher than the Pointer Generator approach. PEGASUS works very well on our dataset with average ROUGE scores achieving 60-70%, which is a very high score for the abstractive summarization task in NLP. The state-of-the-art ROUGE score for popular text summarization datasets are among 30-60% [Zhang et al., 2019]. This attests to the fact that using NLP techniques, it is possible to accurately generate the FIN that is arguably the most important claim among all the claims in a patent application.

Figure 1 shows the probability distributions of the resulting rouge scores when one uses the Pegasus approach on our test set of 31,644 data points.

Based on the obtained probability distributions (i.e., probability density functions) shown in Figure 1, and the expected value (mean) of these distributions, one can observe that our chosen model performs very well on the dataset.

7 Discussion

While the results reported look very promising, our current experiments have the following limitations: 1) We used abstractive summarization to generate the first independent claim which limits the input length. This might be solved by using a combination of extractive summarization, like heuristics, and abstractive summarization; 2) Our results show that our generated claim is mainly a summary of what the inventor writes. However, the claims generated should also establish the novelty of the invention by looking at other patents and/or published papers in the public domain for similar inventions in the same space, instead of only focusing on one single patent application, which requires incorporating external knowledge; 3) Using ROUGE score as the evaluation metric, which focuses on the syntax similarities between the generated claims and the gold standards, might not be sufficient to evaluate the quality of the generated claims. Other aspects, such as semantic similarity, factuality, etc. need to be considered as well. We plan to address these problems in future work.

8 Conclusion

Our results suggest that claim construction and the patent application process can be largely automated in the future with the help of AI, natural language processing, and machine learning. This will have far-reaching consequences such as:

- democratizing the landscape for innovation and inventions, thus enabling small businesses, underrepresented groups, and individual inventors (in addition to big companies) to file for patents in a much more cost-effective manner to own IP;
- expediting the submission and issuing of patent applications dramatically (from 3 or 4 years to less than 1 year), thus making the IP litigation process much more efficient;
- facilitating disruptive changes in the IP litigation process by AI where the machine will be able to do some, if not most, of the legal tasks currently performed by humans.

We hope that our results will stimulate further research into using AI, Natural Language Processing, and Machine Learning for automating the patent application process.

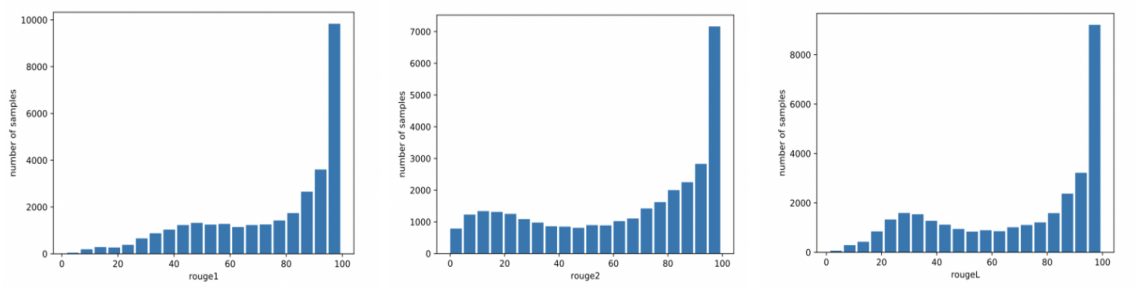

Figure 1: Probability Distribution of F1 scores for ROUGE-1, ROUGE-2, and ROUGE-L.

References

Yiran Chen, Pengfei Liu, Ming Zhong, Zi-Yi Dou, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Cdevalsumm: An empirical study of cross-dataset evaluation for neural summarization systems, 2020.

Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. On the use of arxiv as a dataset. *CoRR*, abs/1905.00075, 2019. URL http://arxiv.org/abs/1905.00075.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679, 2021.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy, July 2019. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies, 2020.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Zhiyong Lu. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database*, 2011, 01 2011. ISSN 1758-0463. doi: 10.1093/database/baq036. URL https://doi.org/10.1093/database/baq036. baq036.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization, 2018.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. *CoRR*, abs/1912.08777, 2019.