Employing Active Learning for Training a DL-Model for Citation Identification in Patent Text*

Farag Saad^{1,*,†}, Hidir Aras^{1,†} and Mark Prince^{2,†}

¹FIZ Karlsruhe - Leibniz Institute for Information Infrastructure, Hermann-von-Helmholtz Platz 1 · 76344 Eggenstein-Leopoldshafen, Germany ²CAS - Chemical Abstracts Service, 2540 Olentangy River Rd, Columbus, OH 43202, USA

Abstract

Citations play an important role in patent analytics. Due to the fact that existing citation lists in patent documents are incomplete, detecting and enhancing them automatically from the patent text has been a user need in patent information retrieval since a while. In this paper, we describe an approach for the identification of citations in patent text using Deep Learning (DL) models. We apply active learning for training and improving of a DL-based named entity recognition (NER) model for this task. The evaluation showed a high accuracy for the focused type of citations, i.e. for the p-c-p (patent cites patent) case.

Keywords

Patent Citations, Named Entity Recognition, Deep learning, Active learning

1. Introduction

Citations play an important role in patent retrieval and analytics. Since the existing citation lists in patent documents are incomplete, it is a long-cherished user wish to automatically determine and complete them from the patent text. Furthermore, citations within a full-text of a patent do not follow well-defined patterns or rules, therefore identifying them with high accuracy is a challenging task. In addition, researching and implementing a suitable approach for utilizing citations from patent text depends on the use case (e.g., search for prior art, linking to corresponding online sources, etc.). Successfully creating such citation lists for patents automatically from patent text will enable users to extend their discovery more efficiently to stated background or adjacent prior art.

In general, patent citations typically come in two types: a patent cites another patent (herein referred to as p-c-p) or a patent cites literature (herein referred to as p-c-l) referring to as NPL (non patent literature) citations. However, in this paper the focus will be on the p-c-p use case where we have designed, implemented and evaluated an approach for patent citation identification based on the chosen p-c-p citation type.

There are two citations patterns types for the p-c-p citation use case, the standard citation pattern type and the non-standard citation pattern type. In the standard citation pattern type, patent applicants tend to use a simple form for referencing other patent publications e.g.,

Corresponding author.

"US20050114951A1, WO 2006122188" etc., while in the nonstandard citation pattern type patent applicants tend to use more complex pattern for citing other patents e.g., "U.S. Pat. Nos. 6,808,085; 6,736,293; 6,732,955; 6,708,846; 6,626,379; 6,626,330; 6,626,328; 6,454,185, United States Provisional Application No.61/914,561, Japanese Unexamined Patent Publication No. 4-187748, US provisional application Serial No 61/640,128" etc. Based on that, we have developed and trained a suitable p-c-p NER DL model for identifying and extracting those types of citation patterns automatically (see Section 3).

In the following, we firstly review shortly the related work in Section 2, followed by a presentation of the proposed approach in Section 3. In Section 4 an empirical evaluation is presented and discussed. Some hints about future work is given in Section 5. A conclusion is given in Section 6.

2. Related Work

Several machine learning approaches have been applied to the problem of extracting data from free text (NER) i.e, citation extraction, among these approaches Support Vector Machines (SVM) e.g., [1], Hidden Markov Models (HMM) e.g., [2], Conditional Random Fields (CRF) e.g., [3]. However, in the past few years, Deep Learning (DL) approaches for the NER task (mainly LSTM = Long Short-Term Memory, CNN = Convolutional Neural Network became dominant as they outperformed the state-of-the-art approaches significantly [4]. In contrast to machine learning approaches, where features are designed and prepared through human effort, deep learning is able to automatically discover hidden features from unlabelled data. The first application for NER using a neural network (NN) was proposed in [5]. In this paper

CEUR Workshop ISSN 1613-0073 Proceedings

⁵th Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech) 2024

[☆] farag.saad@fiz-karlsruhe.de (F. Saad);

hidir.aras@fiz-karlsruhe.de (H. Aras); MPrince@cas.org (M. Prince) © 2024 (copyright for this paper by its authors. Use permitted under Creative Commons Licens Authority of the International (ICC BY 4.0).

the authors considered six standard NLP tasks, among them the NER task where atomic elements in the sentence were labelled into categories such as "PERSON", "COMPANY", or "LOCATION". The authors used feature vectors generated from all words in an unlabelled corpora. A separate feature (orthographic) is included based on the assumption that a capital letter at the beginning of a word is a strong indication that the word is a named entity. The proposed controlled features were later replaced with word embeddings [6] [7]. Word embeddings, which is a representation of word meanings in n-dimensional space, were learned from unlabelled data. A major strength of these approaches is that they allow the design of training algorithms that avoid task-specific engineering and instead rely on large, unlabelled data to discover internal word representations that are useful for the NER task.

3. P-C-P Approach based on Deep Learning

To find suitable training data which can be used to train the p-c-p NER model, we have firstly investigated if there is any publicly available training dataset that we can rely on to build the NER precursor model which will be used to enlarge the training data we have in hand. A precursor model is a type of temporary model which is trained on a small set of training data and will be re-trained further based on a larger training data.

In the following, we give some insight about the publicly available training data as well as the generated training data using the active learning framework.

3.1. Public Training Dataset

We have identified two freely available datasets: the GRO-BID¹ and the manually/expert-created dataset at FIZ Karlsruhe. The GROBID project is specialized on literature citation extraction. However, they have recently done some work related to patent citation extraction. After a pre-processing steps e.g., removing non-English patents, corrupted documents etc., the total number of obtained annotated documents were 130 belonging to three patent authorities: EPO², US³ and PCT (WIPO)⁴. The citations coverage for each patent document is varying, there is some document which contains two citations while some other documents contain up to 375 citations. The domains of focus is life science. The total number of the extracted training paragraphs that hold citations were 487. The FIZ dataset contains 41 annotated patent documents. The citation coverage for each patent varies, between 2 and 66 citations. The domains of focus are life science and technology. The total number of the extracted training paragraphs were 230. Both datasets have a different format, not equivalent to the NER state-of-the art format e.g., *IOB* (Inside-outside-beginning (tagging)) format (See [4]). We have unified the format of both datasets to the standard IOB format and processed the resulting training data (with 717 training paragraphs) to train the precursor p-c-p NER model.

3.2. Generated P-C-P Training Data

For training the precursor p-c-p NER model the freely acquired training data is not sufficient and needs to be improved in quality and quantity. Therefore, we have generated our own training data. To achieve this goal, patent paragraphs that hold many citations are pre-annotated by the p-c-p precursor model first. Then, the patent subject matter experts (SMEs) used the visual annotation user interface of the Prodigy⁵ annotation tool to review and enhance the annotated p-c-p citations. Prodigy is an annotation tool with a easy to use interactive interface that supports by active learning. It is a scriptable tool that allows users to create the annotation themselves, enabling rapid iterations.

From the utilized patent full-text databases PCT (WIPO) and US we have prepared 500 (in total 1000) citation-rich paragraphs belonging to an equally distributes set of patent documents based on their IPC/CPC⁶ classes. In order to use citation-rich paragraphs for training, we have kept only the part of the detailed description (DETD) which holds at least 8 cited patents identified by the precursor p-c-p model. Furthermore, we took into account to have sufficient training instance candidates (in the selected paragraphs) that represent the two types of citation patterns, the standard as well as the non-standard one. The training instance candidates are then reviewed by the SMEs to accept or correct each instance using the Prodigy annotation tool (See Figure 1).

3.3. Model Design and Training

Figure 2 shows the workflow how the final NER model (based on Convolutional Neural Networks (CNNs)) is built for the p-c-p task within the active learning framework. To train the NER model we have utilized the open source framework spaCy⁷. For rapid implementation, we have used the spaCy implementation which is provided by the Prodigy framework. As spaCy offers no pre-

¹https://github.com/kermitt2/grobid/tree/master/grobid-

trainer/resources/dataset

²https://www.epo.org/en

 $^{^{3}} https://www.cas.org/support/training/stnanavist/uspatfull-anavist$

 $^{{}^{4}} https://www.cas.org/support/training/stnanavist/pctfull-anavist$

⁵https://prodi.gy/

⁶https://www.wipo.int/classifications/ipc/en/

⁷https://spacy.io/

REFERENCE 1

prodigy

PROJECT INFO

NTASET INGUAGE ICIPE	en	[DESC0070] Exemplary solid supports are provided in the following patents, published patent applications and other publications,
EWID	which are incorporated in their entirety by reference herein; U.S. Pat. No. 4,336,173 REFERENCE , U.S. Pat. No. 4,459,378 REFERENCE	
DOCDECC		, U.S. Pat. No. 4,654,267 REFERENCE , U.S. Pat. No. 5,234,809 REFERENCE , U.S. Pat. No. 5,705,628 REFERENCE , U.S. Pat. No.
	5,898,071 reference , WO 96/09379 reference , WO 99/58664 reference , WO 02/055727 reference , WO 05/089929 reference	
ITAL	o	, U.S. Pat. No. 5,648,124 REFERENCE , WO 96/37313 REFERENCE , U.S. Pat. No. 6,534,262 REFERENCE , US 2003-0235839 REFERENCE ,
	0%	US 2004-0214175 REFERENCE , US2006-0003357 REFERENCE , U.S. Pat. No. 6,812,341 REFERENCE , U.S. Pat. No. 7,052,840 REFERENCE ,
ISTORY		US 2006-0205004 REFERENCE , WO 03/085091 REFERENCE , US 2002-0106686 REFERENCE , US 2006-0024701 REFERENCE , US 2004-
		0215011 REFERENCE , US 2006-0058519 REFERENCE , US 2005-0239068, US 2005-0106577, US 2005-0106576 REFERENCE , US
		2005-0106589 REFERENCE , US 2005-0106602, US 2005-0136477 REFERENCE , WO 06/036243 REFERENCE , WO 06/036246 REFERENCE
, WO 06/019387 REFERENCE , WO 06/19388 REFERENCE , WO 06/019568 REFERENCE , US 2006-0177836 REFERENCE , WO		
		06/015326 REFERENCE , US2005-0027116 REFERENCE , US 2002-0177698 REFERENCE , U.S. Pat. No. 6,646,118 REFERENCE , U.S. Pat.
		No. 6,429,309 Reference , WO 94/11103 Reference , WO 97/08547 Reference , WO 01/81566 Reference , WO 04/020449
		REFERENCE , U.S. Pat. No. 5,976,426 REFERENCE , US 2005-0181378, US 2005-0196856 REFERENCE , US 2006-0160122 REFERENCE ,
		US 2005-0271553 REFERENCE , DeAngelis et
		Acids Research 23(22):4742-4743 (1995), F 🧹 🗙 🕢 🔁 using a solid-phase" Nucleic Acids
2017-2022 Explosion (Prodigy	v1.11.7)	Research 22(21):4543-4544 (1994), Aviv et messenger RNA by chromatography on

Figure 1: Pre-annotations provided by the p-c-p model and displayed in the Prodigy tool interface.

trained NER model for identifying citations in patents text, we have trained our own DL-based p-c-p NER model utilizing the Prodigy framework (See Figure 2).

The process starts by training a precursor model based on the public dataset, which is then utilized to enlarge the training data iteratively (See Figure 2 (1)). This precursor model was used to filter the acquired US and PCT raw data where we have kept only the part of the Detailed Description (DETD) text which holds at least 8 patent citations (See Figure 2 (2)). We have then pre-processed and prepared the raw data to ensure that it holds sufficient citations. We then loaded part of it into the Prodigy tool along with the integrated precursor model to start the training process for the final NER model in several iterations. The input training instance candidates (in the selected paragraphs), which were annotated by the precursor model, are loaded into the Prodigy tool and presented to the SMEs for reviewing. The SMEs interacted with the pre-annotations and either approved, corrected or added new annotations (See Figure 2 (3)) in the presented paragraphs.

In the first iteration, the SMEs have reviewed 250 preannotated paragraphs for each US and PCT databases. We have used these reviewed pre-annotated paragraphs to re-train the NER model (See Figure 2 (4)) where the model reached F-Score of 85% and hence another iteration was required. To enhance the model performance, We have picked up more raw data and have pre-annotated it again (250 for each US and PCT databases), using the enhanced p-c-p NER model (See Figure 2 (2)). The newly prepared citation-rich paragraphs were reviewed by the SMEs and used to re-train the NER model. If needed, this process will be iteratively repeated and will end when we reach a certain degree of confidence that the final NER model is significantly trained to be applied for the p-c-p citation identification task (See Figure 2 (4)).

4. Evaluation

Once the p-c-p model is sufficiently trained, the citation identification approach is evaluated by processing the evaluation corpus of 245 patents (prepared by SMEs) representing a random collection of patents from the US (128 patents) and PCT (117 patents) patents. To start the p-c-p model evaluation process, all required materials e.g., identified citations of the SMEs evaluation corpus, etc., were handed over to the SMEs.

We have evaluated the p-c-p model based on the *Precision* (*P*) (See equation 1), *Recall* (*R*) (See equation 2), and F1 - Score measures (See equation 3). In order to compute the scores, the False Positive (*FP*), True Positive (*TP*) and False Negative (*FP*) counts are determined first. The *FP* refer to the number of wrongly identified citations by the p-c-p model. The *TP* refer to



2nd - iteration

Nth - iteration

Figure 2: Building the p-c-p ner model and improving it through Experts' interaction

Model training 8

evaluation

Table 1

The p-c-p ner model overall score for precision, recall and F1-Score

DATABASE	Identified citations	FP	TP	FN	Precision	Recall	F1-Score
US	727	17	710	23	0.97	0.96	0.96
PCT	251	11	241	47	0,95	0.83	0.88
Summation	978	28	951	70	0.96	0.89	0.92

the number of correctly identified citations by the p-c-p model. The FN refers to the number of citations that the model fails to identify.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 - Score = \frac{2 * (Precision * Recall)}{Precision + Recall}$$
(3)

As it is shown in Table 1 in total the p-c-p model successfully identified 951 citations out of 978 citations and failed to identify only 28 citations. An explanation of these fails is that in rare cases patent applicants don't cite citations in the right way, for example, consider this citation example "*This is a continuation-in-part of my copending application Ser. No. 43,784 for Catalytic Reformer Process*". Here the patent authority marker (*US, WO, JP*, etc.) is missing so the model has no clue about

which patent authority is meant, therefore, to minimize error rate the model was trained to neglect such incomplete citation.

Even though the p-c-p model obtained a higher evaluation score also for different patent authorities e.g., Finish, Japanese etc. that it was not trained on, it failed to identify some citations that appear in some special context. An effective solution for these failures is to increase the training data to cover more patent authorities. This can be done efficiently using the framework described in Section 3.2.

Generally, the p-c-p model performed very well in most cases and could achieve a high precision of 96%, a high recall of 89%. In addition, we have computed the F1-Score measure to take into account both Precision and Recall measure in order to ultimately measure the accuracy of the model. Despite the fact that the p-c-p model was trained on a very small training dataset consisting of 1717 training paragraphs, the F1-score (using the evaluation corpus prepared by the SMEs) shows that the p-c-p model achieves a certain degree of accuracy and reaches the F1-Score of 92%.

(3)

5. Future Work Directions

The DL-based p-c-p NER model has been trained with a small set of training data (1717 training paragraphs) related to two patent authorities US and PCT. Even though the model obtained a higher evaluation score also for different patent authority data. However, the developed model needs further training and testing to cover more citations belonging to different patent authorities e.g., to cover more citation patterns which might be specialized to some patent authority. Based on our experience so far, a few thousand training paragraphs for each patent authority should be sufficient. To speed up this process, the developed visual active learning approach in this paper can be utilized.

To utilize the extracted citation for further tasks or application such as search, linking patents with a literature knowledge base through citation etc., the extracted citations need to be post-processed. This importantly needed as significant portions of the identified citations were presented in the non-standard citation form e.g., *U.S. Pat. Nos. 5*,*188*,*960*; *5*,*689*,*052*; *5*,*880*,*275*; *5*,*986*,*177*; *7*,*105*,*332*; *7*,*208*,*474*. Hence, after extraction, the individual citations should be normalized accordingly: *US5188960*, *US5689052*, *US5880275*, *etc.* Another example for the normalization is splitting up the identified citation string e.g., *"EP 0 716 884 A2"* into meaningful segments: The patent authority "EP", the patent number "0716884", the patent kind code "A2", and, finally the normalized patent string "EP0716884A2".

To consider a detailed patent citation type such as the filling number of a patent application, the publication of a patent application etc., it is essential to integrate a patent citation specific scheme into the developed approach. For example, if we consider the US patent citation, we noticed that the filing number of a US patent application has a specific format (e.g., No.16/769,261), the publication of a US patent application for a US patent application for a US patent application has a specific format (e.g., US 2005/0114951 A1; starting with a year number) and a US patent has a specific format (e.g., US 6,808,085). Encoding such features into the developed approach will certainly lead to a significant improvement.

6. Conclusion

In this paper, we have developed a DL-based p-c-p NER model to identify citations in the patent fulltext. To realize that, we have designed, implemented and evaluated an active learning framework for patent citation identification employing a DL approach. Furthermore, to train a robust citation identification p-c-p model with high accuracy, we have designed an active learning framework that can be used by patent SMEs to iteratively improve the model performance significantly with less manual effort. In the first iteration, the reviewed pre-annotated paragraphs (250 for each US and PCT databases) have led to a significant model improvement. However, another iteration which involved another 250 pre-annotated paragraphs for each database was required in order to achieve the desired F1-Score of 92%, and to stop the re-training process.

References

- X. Zhang, J. Zou, D. X. Le, G. R. Thoma, A structural svm approach for reference parsing, in: 2010 Ninth International Conference on Machine Learning and Applications, 2010, pp. 479–484. doi:10. 1109/ICMLA.2010.77.
- B. Ojokoh, M. Zhang, J. Tang, A trigram hidden markov model for metadata extraction from heterogeneous references, Information Sciences 181 (2011) 1538–1551. URL: https://www.sciencedirect.com/ science/article/pii/S0020025511000259. doi:https: //doi.org/10.1016/j.ins.2011.01.014.
- [3] P. Lopez, Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications, in: European Conference on Research and Advanced Technology for Digital Libraries, 2009. URL: https://api.semanticscholar.org/ CorpusID:27383212.
- [4] F. Saad, H. Aras, R. Hackl-Sommer, Improving named entity recognition for biomedical and patent data using bi-lstm deep neural network models, in: E. Métais, F. Meziane, H. Horacek, P. Cimiano (Eds.), Natural Language Processing and Information Systems - 25th International Conference on Applications of Natural Language to Information Systems, NLDB 2020, Saarbrücken, Germany, June 24-26, 2020, Proceedings, volume 12089 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 25–36. URL: https://doi.org/10.1007/978-3-030-51310-8_3. doi:10.1007/978-3-030-51310-8_3.
- [5] R. Collobert, J. Weston, A unified architecture for natural language processing: Deep neural networks with multitask learning, in: Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 160–167.
- [6] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. P. Kuksa, Natural language processing (almost) from scratch, Computing Research Repository - CORR abs/1103.0398 (2011).
- [7] E. Parsaeimehr, M. Fartash, J. A. Torkestani, Improving feature extraction using a hybrid of cnn and lstm for entity identification, Neural Process Lett 55 (2023) 5979–5994. doi:https://doi.org/ 10.1007/s11063-022-11122-y.