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

Farag Saad[1] , Hidir Aras[1] and Mark Prince[2]

[1] FIZ Karlsruhe, Karlsruhe, Germany / firstname.lastname@fiz-karlsruhe.de
[2] CAS - Chemical Abstracts Service, Columbus, Ohio, 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.

Patent P-C-P Citation Patterns Types
There are two citations patterns types for the p-c-p citation use case:
1. The standard citation pattern type where patent applicants tend to use a simple form for referencing other patent publications e.g., "US20050114951A1, WO 2006122188"

2. The non-standard citation pattern type where 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.

P-C-P APPROACH BASED ON DEEP LEARNING
Training Data:
- Based on a publicly available training dataset we have built a p-c-p NER precursor model to annotate more raw data extracted from patent documents
- The extracted raw data belongs to the patent full-text databases PCT (WIPO) and US where we have prepared 500 (in total 1000) citation-rich paragraphs belonging to an equally distributes set of patent documents based on their IPC/CPC classes.
- As shown in Figure 1, the patent subject matter experts (SMEs) used the visual annotation user interface of the Prodigy annotation tool to review and enhance the annotated p-c-p raw data

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

Model Design and Training (Active Learning Scenario):
Figure 2 shows the steps to build a final NER model (based on Convolutional Neural Networks (CNNs)) within an active learning framework:
- **Figure 2 (1):** The process starts by training a precursor NER model which is then utilized to enlarge the training data iteratively.
- **Figure 2 (2):** The precursor NER model is used to filter the acquired US and PCT patent raw data i.e., keep part of the raw patent data that holds at least 8 patent citations.
- **Figure 2 (3):** In the first iteration, part of the filtered raw data, which were annotated by the precursor model (pre-annotations), 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 in the presented paragraphs.
- **Figure 2 (4):** The reviewed pre-annotations is used to re-train the NER model. To enhance the model performance, more filtered data is picked up and is pre-annotated. The newly pre-annotation are reviewed by the SMEs and used to re-train further 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.

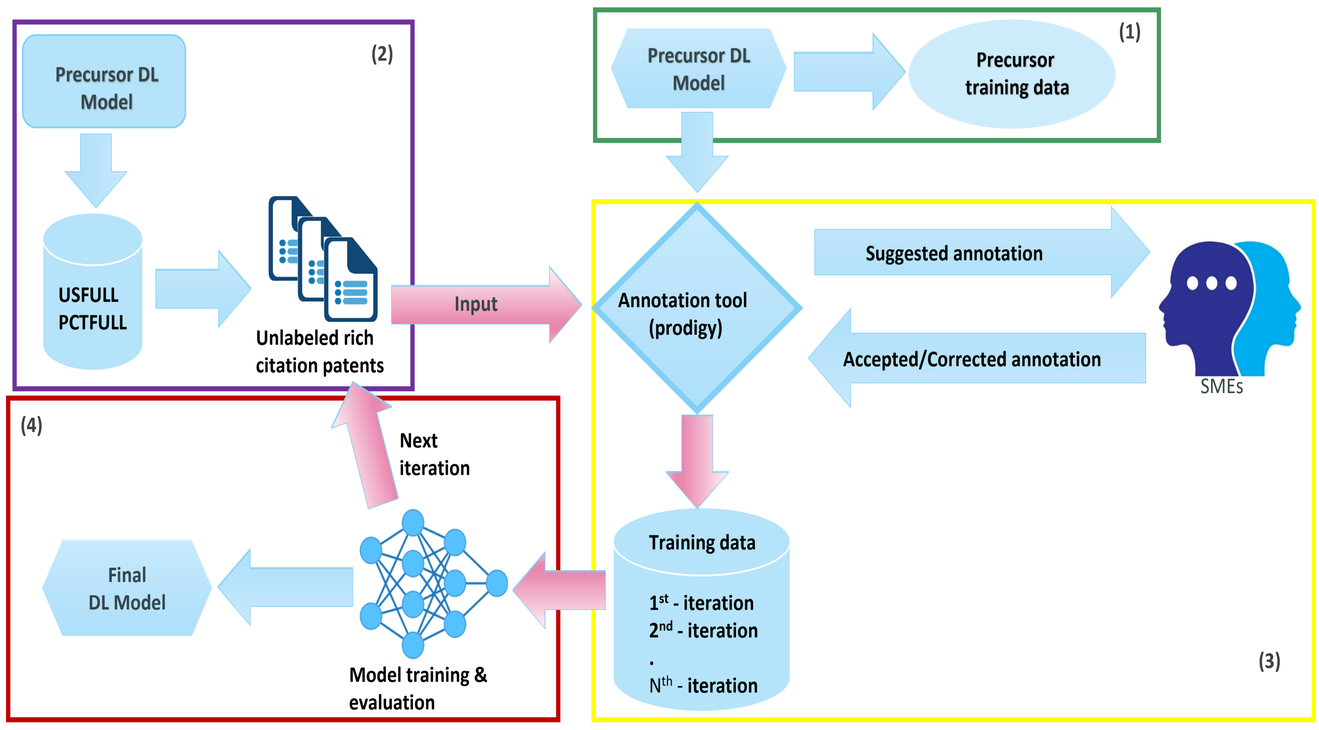

Figure 2 - Building the p-c-p NER model and improving it through Experts' interaction

Experimental Results

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 approach was evaluated based on 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
- 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 and F1-Score of 92%

Future work

- The DL-based p-c-p NER model has been trained with a small set of training data related to two patent authorities US and PCT. However, further training and testing to cover more citations belonging to different patent authorities is required e.g., to cover more citation patterns which might be specialized to some patent authority.
- 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. For example splitting up the identified citation string e.g., ”EP 0716 884 A2” into meaningful segments: The patent authority ”EP”, the patent number ”0716884”, the patent kind code ”A2”, and, the normalized patent string ”EP0716884A2”.

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS