What is Special about Patent Information Extraction?

Liang Chen†
Institute of Scientific and Technical Information of China
Beijing, China P.R.
25565853@qq.com

Shuo Xu
College of Economics and Management
Beijing University of Technology
Beijing, China P.R.
xushuo@bjut.edu.cn

Weijiao Shang
Research Institute of Forestry Policy and Information
Chinese Academy of Forestry
Beijing, China P.R.
shangwj490@163.com

Zheng Wang
Institute of Scientific and Technical Information of China
Beijing, China P.R.
wangz@istic.ac.cn

Chao Wei
Institute of Scientific and Technical Information of China
Beijing, China P.R.
weichaolx@gmail.com

Haiyun Xu
Chengdu Library and Information Center
Chinese Academy of Sciences
Beijing, China P.R.
xuhy@clas.ac.cn

ABSTRACT

Information extraction is the fundamental technique for text-based patent analysis in era of big data. However, the sentences in patent documents are more lengthy and syntactically complicated, and have much more professional terms, which enable the performance of general information-extraction tools to reduce greatly. In this paper, we explore the special characteristics of patent annotation dataset and its impact on information extraction from three aspects: (1) what is the special characteristics of patent annotation dataset? (2) How to choose word embeddings for patent information extraction? (3) Which model is more suitable for your patent annotation dataset? Finally, several suggestions are provided for patent information extraction research in future.

CCS CONCEPTS

CCS→Information systems→Information retrieval→Retrieval tasks and goals→Information extraction

KEYWORDS

Patent information extraction, deep learning, word embedding

1 Introduction

According to the definition of WIPO [1], a patent is an exclusive right granted for an invention, which is a product or a process that provides, in general, a new way of doing something, or offers a new technical solution to a problem. To get a patent, technical information about the invention must be disclosed to the public in a patent application. Patent information can be divided into three types, (1) structured information, which means the bibliographic data in front pages of patent documents, such as applicants, filing date, IPC codes and references, (2) unstructured information, which means the unstructured text sections in patent documents such as abstracts, descriptions and claims. (3) drawings, which include a series of figures to describe the details of the inventions.

For now, data mining for the structured information is much more mature in terms of the methodology and analysis techniques. However, most novel technological information is hidden in the unstructured information of patents. Thus we use information extraction as the vital tool to extract limited kinds of semantic content from text. The main process of information extraction includes NER (Named Entity Recognition), RE (Relation Extraction) and other steps such as entity linking, knowledge base completion. In this paper, we focus on NER and RE according to their significance in information extraction.

Because of the underlying legal purpose of patent documents, patent writers need to define the scope of an invention and need to delimit it from others whilst covering as much variation as possible. As a consequence, patent text is quite different from generic text such as news, encyclopedias in terms of content, genre and vocabulary, and these differences will have a great impact on patent information extraction. In this paper, we explore the special characteristics of patent annotation dataset and its impact on information extraction from three aspects, thus to give suggestions for patent information extraction research in future, (1) what is the special characteristics of patent annotation dataset ? (2) How to choose word embeddings for patent information extraction? (3) Which model is more suitable for your patent annotation dataset?

2 What is special about patent annotation dataset?

This section explores the characteristics of patent annotation dataset via comparative analysis. We collected seven annotation datasets which can be divided into three categories: (1) news corpora consisting of Conll-2003 [2] and NYT-2010 (New York Times corpus) [3], (2) encyclopedia corpora consisting of Wikigold [4] and LIC-2019 (the annotated dataset of 2019 langua-

Table 1 The summary of different annotation datasets

	corpus description	average length of sentence	# of entities per sentence	# of words per entity	# of relations per sentence	entity repetition rate	relation repetition rate	# of unigram entities VS # of ngram entities
CPC-2014(EN)	Patent full-text regarding biology and chemistry	23.3	2.5	1.4	---	5.3	---	74.3:25.7
CGP-2017(EN)	Patent abstract regarding biomed-ical science	21.9	2.4	1.3	0.6	3.7	4.73	80.7:19.3
TFH-2020(EN)	Patent abstract regarding thin film head techniques	30.7	6.1	2.3	4.3	2.8	1.2	24.5 : 75.5
Conll-2003(EN)	Reuters news stories	4.9	1.7	1.5	---	37.7	---	62.4 : 37.6
Wikigold(EN)	Wikipedia	23.0	2.1	1.8	---	5.1	---	49.6 : 50.4
NYTC(EN)	New York Times Corpus	40.6	2.2	1.5	0.4	13.5	8.0	55.9: 44.1
LIC-2019(CN)	search results of Baidu Search as well as Baidu Zhidao	---	3.0	---	2.1	2.5	1.3	---

-ge and Intelligence Challenge) [5], (3) patent corpora consisting of CPC-2014 (Chemical Patent Corpus) [6], CGP-2017 (The CEMP and GPRO Patents Tracks) [7], TFH-2020 (thin film head annotated dataset) [8].

There are 7 indicators employed to analyze these datasets, which are average sentence length, average count of entities per sentence, average count of relations per sentence, average count of words per entity, entity repetition rate[1], relation repetition rate[2] and the ratio of unigram entities count to n-gram entities count. While calculating the indicators, there are two points worth mentioning: (1) in CGP-2017, Conll-2003 and Wikigold only entities are annotated without semantic relationships, and (2) all datasets are in English except LIC-2019 which is in Chinese. As a consequence, some indicators cannot be calculated for certain datasets, such as average count of relations per sentence for CGP-2017, Conll-2003 and Wikigold, and average sentence length, average count of words per entity, and the ratio of unigram entities count to n-gram entities count for LIC-2019. The indicators are shown in Table 1.

From the statistics above, we can easily find the following facts:

(1) For average sentence length, there is no significant difference between patent text and generic text;

(2) For average count of entities per sentence, patent text is larger than generic text;

(3) There exists significant distinguish between patent datasets from different domains. This distinguish is two-fold: firstly, it comes from the different characteristics of patent text in different domains, e.g., the average sentence length of patents in thin-film magnetic head is larger than that of biology and chemistry; secondly, it comes from the different concerns of the annotators from different domains, e.g., for TFH-2020 dataset, there are 17 types of entities concerned, as to CGP-2017 dataset, only 3 types of entities including chemical, gene-n, gene-y are concerned.

3 What is special about patent word embeddings?

After long-term development, information extraction methods have formed a large family, in which deep learning is the representative of state-of-the-art technologies. Word embedding is the foundation of deep learning for information extraction, it refers to a class of techniques where each word is mapped to one real-valued vector in a predefined vector space and the vector values are learned in a way that resembles a neural network.

There are two ways to obtain word embeddings (1) by training on a corpus via word embedding algorithm, such as Skip-gram [9], CBOW [9] and the like; (2) by directly downloading a pre-trained word embedding file from the Internet, like GloVe [10]. Risch and Krestel [11] suggested obtaining word embeddings by training specifically on patent documents in all fields for improving semantic representation of patent language. In fact, such suggestion is based on automatic classification for patents in all fields, which is quite different from information extraction from patents in specific domain. In order to explore which word embedding is preferable in patent information extraction, four types of word embedding with the same dimensions of 100 are prepared as follows:

(1) word embeddings of GloVe provided by Stanford NLP group. According to the different training corpora, there are four release versions of GloVe [10]. We choose the one trained on *Wikipedia 2014* and *Gigaword 5* as it provides word embeddings of 100 dimensions. In fact, the version trained on *Twitter* also has word embeddings of 100 dimensions. But since our training corpus does not follow the patterns in such short texts as Twitter;

(2) word embeddings provided by Risch and Krestel [11], which are trained with the full-text of 5.4 million patents granted from USPTO during 1976 to 2016. Risch and Krestel released three versions of word embeddings with 100/200/300 dimensions. The 100 dimensions version is chosen and referred to it as USPTO-5M;

(3) word embeddings trained with a corpus of 1,010 patents mentioned in this paper but with their full-text (abstract, claims and description), these word embeddings are referred as TFH-1010;

(4) word embeddings trained with the abstract of 46,302 patents regarding magnetic head in hard disk drive, these word embeddings are referred as MH-46K.

On the basis of these word embeddings, we ran BiLSTM-CRF of NER and BiGRU-HAN of RE as shown in Table 2 and Table 3. Even the results produced by these four types of word embedding are almost the same, MH-46K and TFH-1010 still slightly outperform the other word embeddings.

However, as Risch and Krestel [11] reported, the performance improvement is observed in term of micro-average precision when replacing Wikipedia word embeddings with USPTO-5M word embeddings. In our opinion, the main reason may lie in the huge difference between automatic classification for patents in all fields and the information extraction from patents in a specific domain. To say it in another way, when one confronts a task in a specific domain, the word embeddings trained on the same domain corpus should be preferred to.

4 What is special about methods of patent information extraction?

There are mainly 2 manners of information extraction, which are pipelined manner and joint manner. As shown in Fig.1, pipelined manner extracts the entities first and then recognizes their relations, this separated framework makes the information extraction task easy to deal with, and each component can be more flexible. Differently, joint framework is to extract entities together with relations using a single model. Even Zheng et al [12] claimed that the joint framework is capable of integrating the information of entities and relations, thus to improve NER and RE performance in a mutually reinforcing way, as to patent information extraction, one can benefit most from joint manner is

[1] the ratio of all entities to the deduplicated entities

[2] the ratio of all semantic triplets to the deduplicated semantic triplets

Table 2 The summary of NER results for different word embeddings

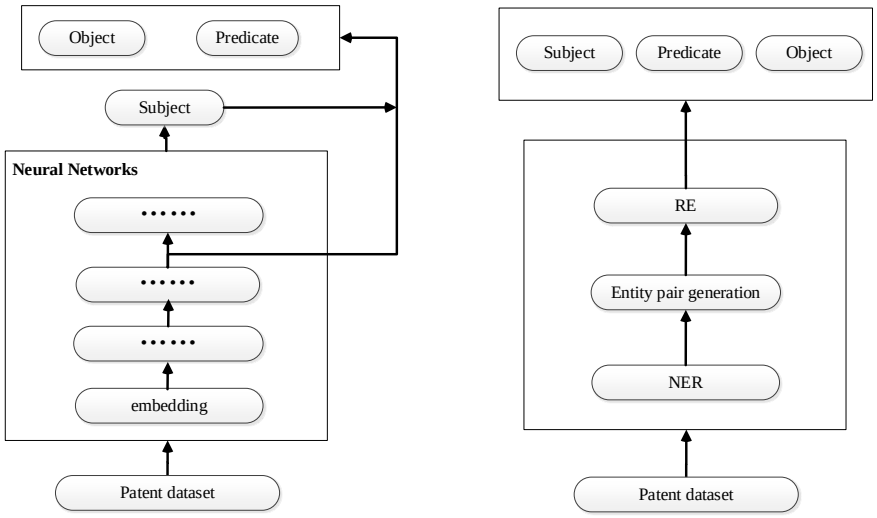
	micro-average			macro-average			weighted-average		
	Precision (%)	**Recall (%)**	**F1 (%)**	**Precision(%)**	**Recall (%)**	**F1 (%)**	**Precision (%)**	**Recall (%)**	**F1 (%)**
GloVe	77.2	77.2	77.2	66.7	**56.0**	**60.9**	**78.6**	77.2	77.9
USPTO-5M	77.1	77.1	77.1	65.1	53.0	58.4	77.9	77.1	77.5
TFH-1010	77.3	77.3	77.3	**67.2**	54.2	60.0	79.1	77.3	78.2
MH-46K	**78.0**	**78.0**	**78.0**	63.9	54.2	58.6	78.5	**78.0**	**78.2**

Table 3 The summary of RE results for different word embeddings

	micro-average			macro-average			weighted-average		
	Precision (%)	**Recall (%)**	**F1 (%)**	**Precision (%)**	**Recall (%)**	**F1 (%)**	**Precision (%)**	**Recall (%)**	**F1 (%)**
GloVe	88.9	88.9	88.9	35.6	28.8	30.0	89.4	88.9	89.0
USPTO-5M	86.9	86.9	86.9	30.8	35.1	31.3	**89.8**	86.9	88.1
TFH-1010	**89.1**	**89.1**	**89.1**	**34.2**	32.1	**32.0**	89.7	**89.1**	**89.3**
MH-46K	87.9	87.9	87.9	31.6	**34.2**	31.6	89.7	87.9	88.6

Table 4 The overall evaluation for different manners of patent information extraction

	micro-average			macro-average			weighted-average		
	Precision(%)	**Recall (%)**	**F1 (%)**	**Precision (%)**	**Recall (%)**	**F1 (%)**	**Precision (%)**	**Recall (%)**	**F1 (%)**
BiGRU-HAN with *no relation*	87.9	87.9	87.9	31.6	34.2	31.6	89.7	87.9	88.6
BiGRU-HAN without *no relation*	41.5	41.5	41.5	27.3.	30.3	27.5	32.3	41.5	36.3
Hybrid Structure of Pointer and Tagging	4.2	4.2	4.2	14.4	2.3	3.7	41.6	4.2	7.6



(a) pipelined method (b) joint learning method

Fig.1 Two manners of information extraction

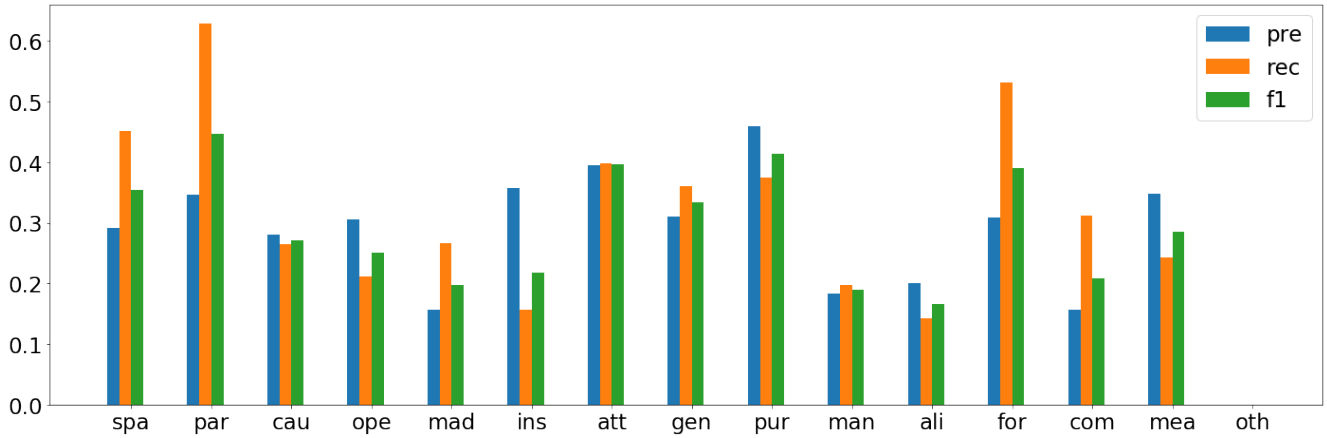


Fig.2 Result of pipelined method for information extraction

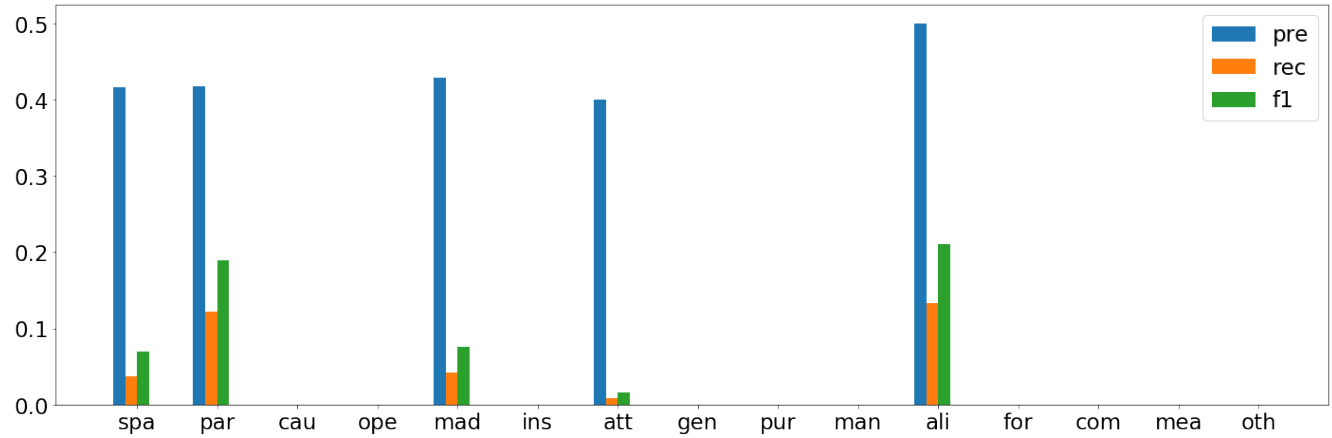


Fig.3 Result of joint method for information extraction

to get rid of entity pair generation which would produce amount of noise data caused by combinatorial explosion.

It seems that joint manner is a better solution for patent information extraction, so what is the actual situation?

To demonstrate the difference of the two manners, a comparative analysis is conducted based on TFH-2020 dataset,and BiLSTM-CRF [13] + BiGRU-HAN [14] and Hybrid Structure of Pointer and Tagging [15] are taken as respective baselines. As TFH-2020 contains much more entities per sentence than generic text, after entity pair generation in pipelined manner, the proportion of no relation is much larger than that of generic text, so for comparative analysis, there are two results provided consisting of BiGRU-HAN with *no relation* and without *no relation*. The final results are shown in Table 4 and Fig.2, Fig.3.

After 40 epochs of training, the results of *Hybrid Structure of Pointer and Tagging* on the test dataset is shown in the third row of Table 6, which is much worse than that of BiGRU-HAN without *no relation*. This is inconsistent with our observation of the information extraction competition in LIC 2019 [5], in which the joint manner outperform the pipelined one by a large margin.

In our opinion, there are mainly two reasons for this situation, (1) as same as pipelined method, the performance of joint model is severely affected by the number of entities in a sentence. (2) The requirement of the joint model on the size of training set is much higher than that of the pipelined model. For a close examination, we take the LIC-2019 dataset as an example to explore how the size of training set affects the joint model's performance, and the result is shown in Fig.4.

We can see the performance of the joint model enhances rapidly as the size of training set increases from 1000 to 50000, after that it is into a stable state around weighted-average precision/recall/F1-value of 0.78/0.51/0.63. It is obvious that for patent datasets with more professional items, the joint model needs a larger training set to achieve such performance. Unfortunately, this is always not feasible for TFH-2020 dataset in this case, pipelined method is still the first choice for patent information extraction.

5 Conclusion

Based on the discussion above, we can conclude that the special characteristics of patent text are two-fold, (1) the special

characteristics of patent text compared to that of generic text, such as entity repetition rate, average number of entities per sentence. (2) The special characteristics of patent text in one domain compared to that of in another domain, such as the average sentence length, the average number of entities and relationships per sentence.

As a consequence, when preparing patent information extraction, both of these characteristics should be taken into consideration. In detail, when one confronts patent information extraction in a specific domain, the word embeddings trained on the same domain corpus should be preferred to. As to choose information extraction model, the scale of training set and the actual performance of the model should be considered, for example, if the size of the training set is less than 10,000, such as TFH-2020 dataset, the pipelined method is preferable.

In addition, there are also big challenges in patent information extraction, including (1) How to generate large-scale patent annotation dataset in an efficient and low-cost way? (2) How to improve patent information extraction via leveraging the special characteristics of patent text? As to the second challenge, we have done some research and achieved remarkable improvement, which indicates that the patent information extraction is quite a promising direction.

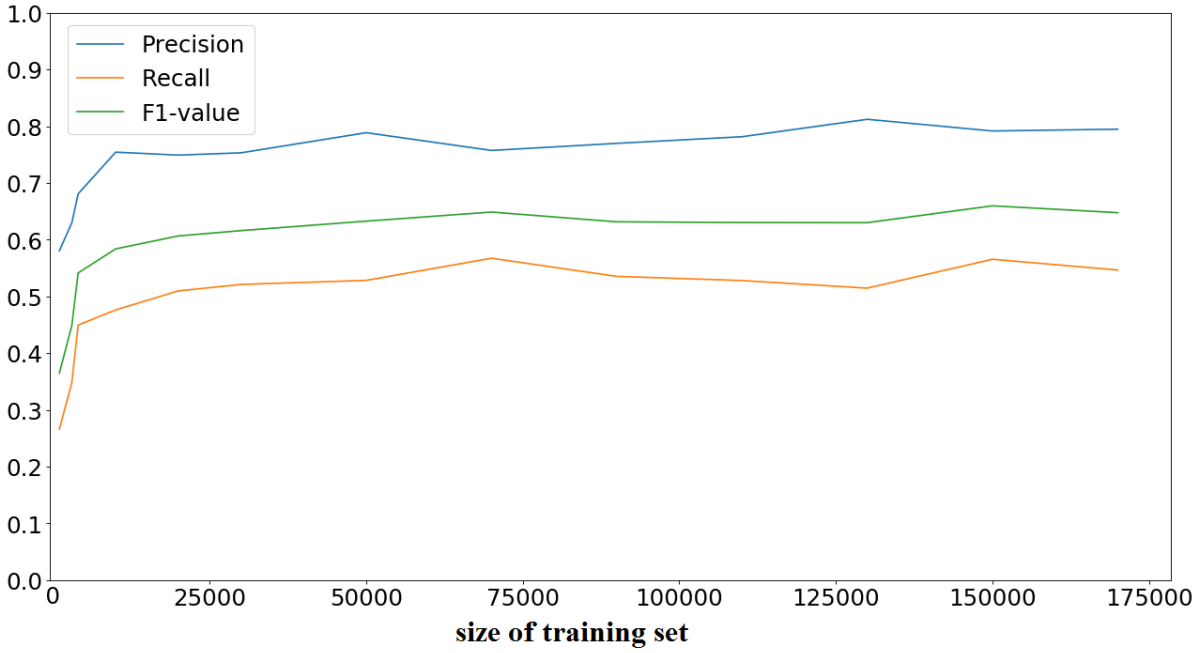


Fig.4 The performance of joint model for training set of different size

ACKNOWLEDGMENTS

This research received the financial support from National Natural Science Foundation of China under grant number 71704169, and Social Science Foundation of Beijing Municipality under grant number 17GLB074, respectively. Our gratitude also goes to the anonymous reviewers for their valuable suggestions and comments.

REFERENCES

[1] What is a patent? https://www.wipo.int/patents/en

[2] Sang E. F. T. K., De Meulder F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. arXiv preprint arXiv:cs/03060-50.

[3] Riedel S., Yao L., McCallum A. (2010) Modeling Relations and Their Mentions without Labeled Text. In: Balcázar J.L., Bonchi F., Gionis A., Sebag M. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2010. Lecture Notes in Computer Science, vol 6323. Springer, Berlin, Heidelberg

[4] Balasuriya D., Ringland N., Nothman J., Murphy T., Curran J. R. (2009). Named Entity Recognition inWikipedia, Proceedings of the 2009 Workshop on the People's Web Meets NLP, ACL-IJCNLP 2009, pages 10–18.

[5] Wu, H.(2019). Report of 2019 language & Intelligence technique evaluation. Baidu Corporation. http://tcci.ccf.org.cn/summit/2019/dlinfo/1101-wh.pdf

[6] Akhondi, S. A., Klenner, A. G., Tyrchan, C., Manchala, A. K., Boppana, K., Lowe, D., Zimmermann, M., Jagarlapudi, S. A. R. P., Sayle, R., Kors, J., & Muresan, S. (2014). Annotated Chemical Patent Corpus: A Gold Standard for Text Mining. PLoS ONE, 9(9), 1-8.

[7] Pérez-Pérez, M., Pérez-Rodríguez, G., Vazquez, M., Fdez-Riverola, F., Oyarzabal, J., Oyarzabal, J., Valencia, A., Lourenço, A., & Krallinger, M. (2017). Evaluation of Chemical and Gene/Protein Entity Recognition Systems at BioCreative V.5: The CEMP and GPRO Patents Tracks. In Proceedings of the BioCreative V.5 Challenge Evaluation Workshop, 11-18.
[8] Chen, L., Xu, S., Zhu, L., Zhang, J., Lei,X., & Yang, G. (2020). A deep learning based method for extracting semantic information from patent documents. Scientometrics.
[9] Mikolov, T., Chen, K., Corrado G.,& Dean, J.(2013). Efficient estimation of word representations in vector Space. arXiv preprint arXiv: 1301.3781.
[10] Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (pp. 1532-1543).
[11] Risch, J., & Krestel, R. (2019). Domain-specific word embeddings for patent classification. Data Technologies and Applications, 53(1), 108–122.
[12] Zheng S.C., Wang F., Bao H.Y., Hao Y.X, Zhou P., Xu B. (2017). Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme. arXiv preprint arXiv:1706.05075
[13] Huang, Z., Xu, W., &Yu K. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.
[14] Han,X., Gao,T., Yao,Y., Ye,D., Liu,Z., Sun, M.(2019). OpenNRE: An Open and Extensible Toolkit for Neural Relation Extraction. arXiv preprint arXiv: 1301.3781
[15] Su, J.L. (2019). Hybrid Structure of Pointer and Tagging for Relation Extraction: A Baseline. https://github.com/bojone/kg-2019-baseline