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 specialty of patent text enables the performance of general information-extraction methods to reduce noticeably. To solve this problem, an in-depth exploration has to be done for clarify the particularity in patent information extraction, thus to point out the direction for further research. In this paper, we discuss the particularity of patent information extraction in three aspects: (1) what is the special about labeled patent dataset? (2) What is special about word embeddings in patent information extraction? (3) What kind of method is more suitable for patent information extraction?

CCS CONCEPTS

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

KEYWORDS

Patent information extraction, deep learning, word embedding

1 Introduction

As an important source of technical intelligence, patents cover more than 90% latest technical information of the world, of which 80% would not be published in other forms [1]. There are two traditional ways to obtain technical intelligence, either by analyzing structured data with bibliometric methods or by experts reading patent texts. However, with the rapid growth of patent documents, the second way is facing more and more challenges.

Information extraction is an important technology for machine understanding text, which aims to extract structured data from free text to eliminate ambiguous problem inherent in free texts. In recent years, with the tremendous advances of machine learning technology, especially the rise of deep learning, the research in information extraction has made great progress. However, the particularity of patent text enables the performance of general information-extraction tools to reduce greatly. Therefore, it is necessary to deeply explore patent information extraction, and provide ideas for subsequent research.

Information extraction is a big topic, but it is far from being well explored in IP (Intelligence Property) field. To the best of our knowledge, there are only three labeled datasets publicly available in the literature which contain annotations of named entity and semantic relation. Therefore, we choose NER (Named Entity Recognition) and RE (Relation Extraction) for discussion in three aspects as follows.

2 What is special about labeled patent dataset?

Since supervised learning methods represent state-of-the-art techniques in information extraction, it's necessary to clarify the particularity of labeled patent dataset for further improvement of information extraction in IP. To this end, a comparative analysis is conducted which contains seven labeled datasets of three categories: (1) news corpus consisting of Conll-2003 [2] and NYT-2010 (New York Times corpus) [3], (2) encyclopedia corpus consisting of Wikigold [4] and LIC-2019 (the annotated dataset of 2019 ;Language and Intelligence Challenge) [5], (3) patent corpus 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 two parts contained in each labeled dataset, (1) an information schema to define label types, (2) a dataset consisting of labeled texts. Let's take TFH-2020 as an example, the schema of named entities and semantic relations are shown in Table 1 and 2 of the Appendix, and the labeled text is shown in Fig. 1.

In order to analyze these datasets, 8 indicators are proposed as shown in Table 3 of the Appendix. It is worth noting that, (1) in CGP-2017, Conll-2003 and Wikigold, only entities are annotated but semantic relations, (2) all datasets are in English except

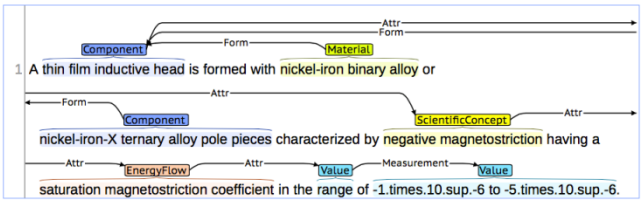

Fig. 1 The labeled text in TFH-2020

LIC-2019, which is in Chinese. As a consequence, some indicators are not calculated for certain datasets. The final result is shown in Table 4 of the Appendix.

From the statistics in Table 4, we can find the following facts:

(1) In terms of *average sentence length*, there is no clear distinction between patent text and generic text;

(2) In terms of *count of entities per sentence*, there are more entities in a single sentence of patent text than that of generic text;

(3) As to rest indicators, TFH-2020 shows clear distinctions from the other patent datasets and all generic datasets.

In summary, there exist significant distinctions not only between patent text and generic text, but also between patent text from different technical domains. In our opinion, the later distinctions are two-fold: firstly, they come from the unique characteristics of different technical domains, i.e., there are plenty of sequences and chemical structures mentioned in describing innovations in *chemical and biotechnology* (Hunt et al., 2012), while the most frequent entities in the field of *hard disk drive* are of components, location and function (Chen et al., 2020), as to describe inventions with different materials and mechanisms, the patents of different domains follow different writing styles; secondly, they come from the concerns of experts from different domains, e.g., for TFH-2020 dataset, there are 17 types of entities designed, as to CGP-2017 dataset, only 3 types of entities including chemical, gene-n, gene-y are concerned while other types of entities are out of concern.

3 What is special about patent word embeddings?

So far deep learning techniques have achieved state-of-the-art performance in information extraction. As the foundation of deep learning techniques in NLP, word embedding refers to a class of techniques where words or phrases from the vocabulary are mapped to vector*s* of real numbers.

There are two ways to obtain word embeddings, (1) by training on a corpus via word embedding algorithm, such as Skip-gram [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 basis of these word embeddings, two deep-learning models, BiLSTM-CRF and BiGRU-HAN, are respectively used for entity identification and semantic relation extraction. Specifically speaking, BiLSTM-CRF (Fig. 2) takes sentences as input and represents every word in a word embedding format, during training procedure these word embeddings pass through the layers within BiLSTM-CRF and output the prediction of named entities in the sentence; The basic idea of BiGRU-HAN (Fig. 3) is to recognize the occurrence pattern of different semantic relations by a recurrent neural network named BiGRU, and then leverages a hierarchical attention mechanism consisting of a word-level attention layer and a sentence-level attention layer to further improve the model's prediction accuracy.

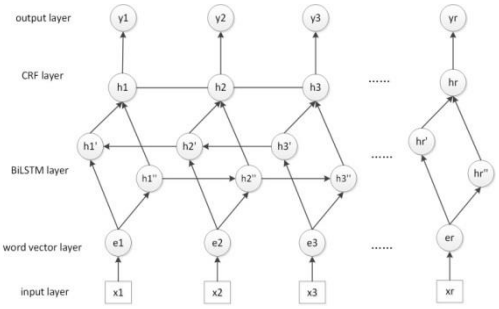

Fig. 2 The structure of BiLSTM-CRF model.

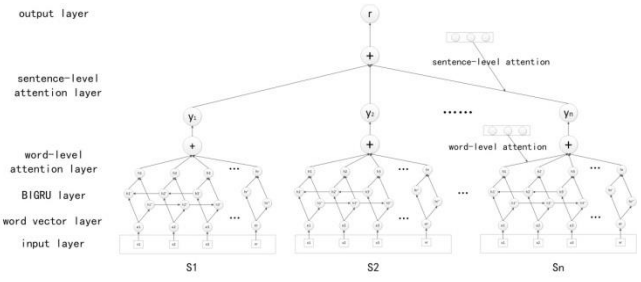

Fig. 3 The structure of BiGRU-HAN model.

From Table 5 and Table 6, we can see the results produced by these four types of word embedding are almost the same. However, Risch and Krestel [11] observed a considerable improvement when replacing Wikipedia word embeddings with USPTO-5M word embeddings in patent classification task. In our

opinion, the main reason lies 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?

As far as supervised learning method is concerned, there are mainly 2 ways for information extraction, namely pipeline way and joint way shown in Fig.4. The former extracts the entities first, and then identifies the relationship between them. The separated strategy makes the information extraction task easy to deal with, and each component can be more flexible; differently, the latter uses a single model to extract entities and relations at one time. Even Zheng et al [12] claimed joint method is capable of integrating the information of entities and relations, thus to improve NER and RE performance in a mutually reinforcing way, in our opinion, the biggest advantage it brings is the elimination of entity pair generation which would produce a large number of entity pairs with *no relation* type shown in Fig. 5.

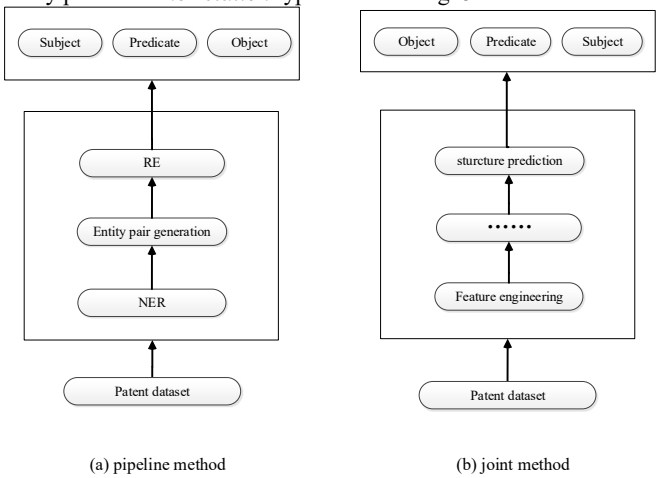

Fig.4 Two patterns of information extraction

Joint method seems to be a better solution to extract patent information, so what is the actual situation?

To verify this, we prepare a pipeline baseline and a joint baseline, namely BiLSTM-CRF [13] &BiGRU-HAN [14] and Hybrid Structure of Pointer and Tagging [15] (Fig. 6) for an experiment on TFH-2020 dataset. Since the proportion of *no relations* in TFH-2020 is much larger than that of generic text after entity pair generation, two set of results are provided by pipeline baseline including with *no relations* and without *no relations*, which are shown in 1st and 2nd rows of Table 7, and the result of *Hybrid Structure of Pointer and Tagging* is shown in 3nd row. In order to highlight the performance of the two baselines on different types of relation, the precision, recall, and F1-value for each type of relation are shown in Fig. 7 and 8 with each type of relation denoted by its first 3 letters (cf. Table 2)

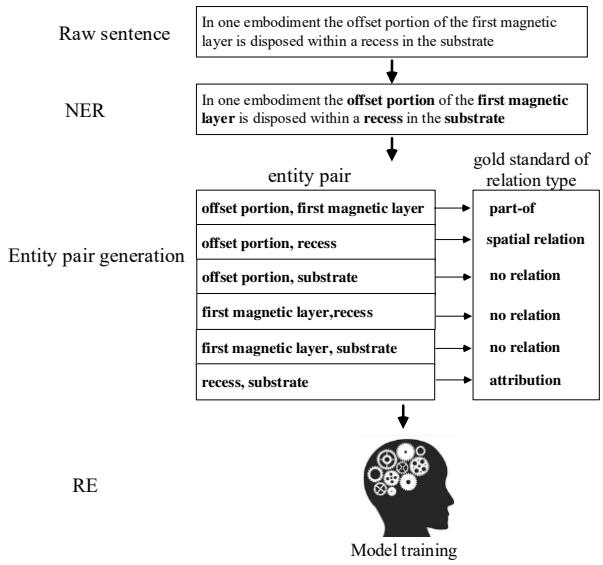

Fig. 5 The procedure of information extraction in pipeline way

As can be seen, the experimental results in this paper contradict the observation from information extraction competition in LIC 2019 [5], where joint methods outperformed pipeline methods by a large margin. In our opinion, there are two reasons behind, (1) as same as pipeline method, the performance of joint method is severely affected by the number of entities in sentences; (2) the requirement of joint model for training set size is much higher than that of pipeline model. To verify the 2nd reason, we take the LIC-2019 dataset as an example to demonstrate how the size of the training set affects the performance of *Hybrid Structure of Pointer and Tagging*.

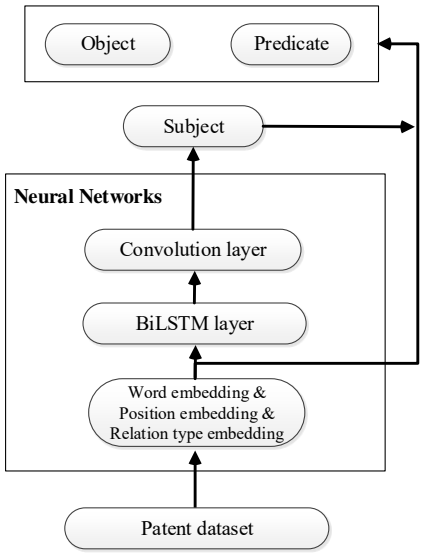

Fig. 6 The structure of Hybrid Structure of Pointer and Tagging

Table 5 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 6 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 7 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

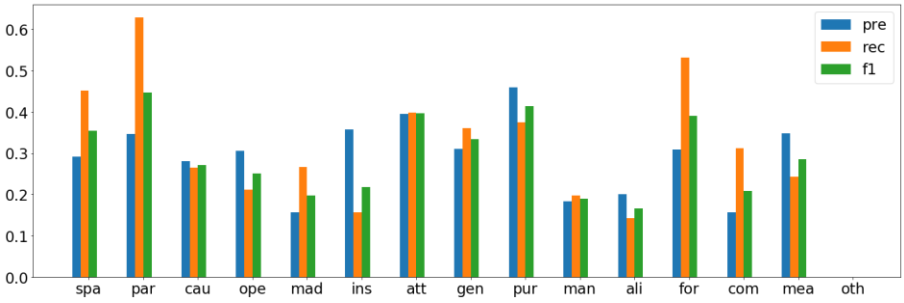

Fig.7 Result of pipeline method for information extraction

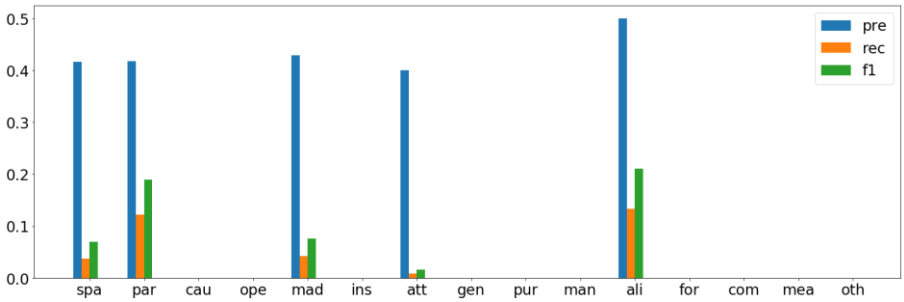

Fig.8 Result of joint method for information extraction

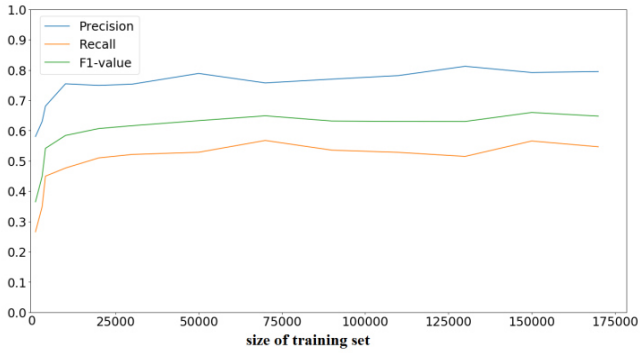

Fig.9 The performance of joint model with different size of training set

As shown in Fig. 9, as the size of the training set increases from 1000 to 50000, the performance of *Hybrid Structure of Pointer and Tagging* increases rapidly, and then it enters a stable state near 0.78/0.51/0.63 in terms of weighted-average precision /recall /F1-value.

5 Conclusions

In this paper, we discuss the particularity in patent information extraction in three aspects:

(1) **Labeled dataset** through comparative analysis, it is found that there are differences not only between labeled patent datasets and labeled generic datasets, but also between labeled patent datasets from different technical fields, which means patent information extraction is a domain-specific task, and a series of processing steps should be customized from feature engineering to model building for better performance.

(2) **Word embedding** word embedding is the foundation of deep learning methods in information extraction. Although Risch et al. suggested obtaining word embeddings by training specifically on patent documents in all fields to improve semantic representation of patent language, experiment shows when one confronts a task in a specific domain, the word embeddings trained on the same domain corpus should be preferred to.

(3) **Organization of sub-tasks in information extraction** although joint method achieves state-of-the-art performance in information extraction, this excellent performance comes at the expense of large labeled dataset. When the dataset is limited, one should take a series of factors, such as model characteristics, computing resources, actual performance and so on into consideration, and then choose an optimal method.

We realize some conclusions in this paper are obtained only considering a few sample data considering simple metrics. However, given the scarcity of patent labeled dataset in information extraction, this is what we can get so far with data support. In the future, we hope more people participate in construction of patent labeled datasets and research of patent information extraction, not only because lack of labeled datasets, but also because there are valuable tasks waiting for us to explore, such as how to generate large-scale patent annotation dataset with low cost? Or how to use the particularity of patent text to improve the performance of information extraction in patent text?

ACKNOWLEDGMENTS

This research received the financial support from National Natural Science Foundation of China under grant number 71704169, National Key Research and Development Program of China under grant number 2019YFA0707202, 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] Zha, X., & Chen, M. (2010). Study on early warning of competitive technical intelligence based on the patent map. Journal of Computers, 5(2). doi:10.4304/jcp.5.2.274-281

[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. doi:10.1007/s11192-020-03634-y.

[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

Appendix:

Table 1 The specification of entity types.

type	comment	example
physical flow	substance that flows freely	The **etchant solution** has a suitable solvent additive such as glycerol or methyl cellulose
information flow	information data	A camera using a film having a magnetic surface for recording **magnetic data** thereon
energy flow	entity relevant to energy	Conductor is utilized for producing **writing flux** in magnetic yoke
measurement	method of measuring something	The curing step takes place at the substrate **temperature** less than 200.degree
value	numerical amount	The curing step takes place at the substrate temperature less than **200.degree**
location	place or position	The legs are thinner near the pole tip than in the **back gap region**
state	particular condition at a specific time	The MR elements are biased to operate in a **magnetically unsaturated mode**
effect	change caused an innovation	Magnetic disk system permits **accurate alignment** of magnetic head with spaced tracks
function	manufacturing technique or activity	A magnetic head having **highly efficient write and read functions** is thereby obtained
shape	the external form or outline of something	**Recess** is filled with non-magnetic material such as glass
component	a part or element of a machine	A pole face of **yoke** is adjacent edge of element remote from surface
attribution	a quality or feature of something	A **pole face** of yoke is adjacent edge of element remote from surface
consequence	The result caused by something or activity	This prevents the slider substrate from **electrostatic damage**
system	a set of things working together as a whole	A **digital recording system** utilizing a magnetoresistive transducer in a magnetic recording head
material	the matter from which a thing is made	Interlayer may comprise material such as **Ta**
scientific concept	terminology used in scientific theory	**Peak intensity ratio** represents an amount hydrophilic radical
other	Not belongs to the above entity types	**Pressure distribution** across air bearing surface is substantially symmetrical side

Table 2 The specification of relation types.

type	comment	example
spatial relation	specify how one entity is located in relation to others	**Gap spacer material** is then deposited on the **film knife-edge**
part-of	the ownership between two entities	a **magnetic head** has a **magnetoresistive element**
causative relation	one entity operates as a cause of the other entity	**Pressure pad** carried another **arm** of spring urges film into contact with head
operation	specify the relation between an activity and its object	**Heat treatment** improves the (100) **orientation**
made-of	one entity is the material for making the other entity	The thin film head includes a **substrate** of **electrically insulative material**
instance-of	the relation between a class and its instance	At least one of the **magnetic layer** is a **free layer**
attribution	one entity is an attribution of the other entity	The **thin film** has very high **heat resistance** of remaining stable at 700.degree
generating	one entity generates another entity	**Buffer layer resistor** create **impedance** that noise introduced to head from disk of drive
purpose	relation between reason/result	**conductor** is utilized for producing **writing flux** in magnetic yoke
in-manner-of	do something in certain way	The **linear array** is angled at a **skew angle**
alias	one entity is also known under another entity's name	The bias structure includes an **antiferromagnetic layer AFM**
formation	an entity acts as a role of the other entity	**Windings** are joined at end to form **center tapped winding**
comparison	compare one entity to the other	**First end** is closer to recording media use than **second end**
measurement	one entity acts as a way to measure the other entity	This provides a relative **permeance** of at least **1000**
other	not belongs to the above types	Then, **MR resistance estimate** during polishing step is calculated from **S valu e** and K value

Table 3 The specification of indicators for comparative analysis

indicator	formula	comment	memo
average length of sentence	$L_{avg} = \frac{\sum_i^N L_i}{N}$	*N* indicates the number of sentences, L_i indicates the length of the *i*-th sentence	Calculate how many words are included in an sentence on average
# of entities per sentence	$SE_{avg} = \frac{\sum_i^N SE_i}{N}$	*N* is the same as above, SE_iindicates the number of entities in the *i*-th sentence	Calculate how many entities are included in an sentence on average
# of words per entity	$EW_{avg} = \frac{\sum_i^{NE} EW_i}{NE}$	*NE* indicates the number of entities in sentences, EW_i indicates the number of words in the *i*-th entity	Calculate how many words are included in an entity on average
# of relations per sentence	$SR_{avg} = \frac{\sum_i^N SR_i}{N}$	*N* is the same as above, SR_iindicates the number of relation mentions in the *i*-th sentence	Calculate how many relation mentions are included in an sentence on average
entity repetition rate	$ER = \frac{NE}{NE\_distinct}$	NE is the same as above, *NE_distinct* indicates the number of entities after deduplication	Calculate how many times an entity can appear in the corpus on average
relation repetition rate	$RR = \frac{RE}{RE\_distinct}$	RE indicates the number of relation mentions in sentences, *RE_distinct* indicates the number of relation mentions after deduplication	Calculate how many times an relation mention can appear in the corpus on average
percentage of ngram entities	$EP_{ngram} = \frac{NE_{ngram}}{NE}$	NE is the same as above, NE_{ngram} indicates the number of multi-word entities, namely ngram entities in sentences	Measure the proportion of phrase-type entities in all entities
entity association rate	$EA = \frac{100 * \sum_i^{NE\_distinct} NE\_associated_i}{NE\_distinct^2}$	$NE\_distinct$ is same as above, $NE\_associated_i$ indicates the number of deduplicated entities that have common word(s) with the *i*-th entity	Measure the connection between entities by co-word mechanism, i.e., thin film head and Ferrite head are connected as they have a common word head

Table 4 The summary of different labeled 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**	**percentage of ngram entities (%)**	**entity association rate**
CPC-2014(EN)	Patent full-text regarding biology and chemistry	23.3	2.5	1.4	---	5.3	---	25.7	1.6
CGP-2017(EN)	Patent abstract regarding biomed-ical science	21.9	2.4	1.3	0.6	3.7	4.73	19.3	0.4
TFH-2020(EN)	Patent abstract regarding thin film head techniques	30.7	**6.1**	**2.3**	**4.3**	2.8	1.2	**75.5**	**3.4**
Conll-2003(EN)	Reuters news stories	14.6	1.7	1.5	---	33.3	---	37.6	0.1
Wikigold(EN)	Wikipedia	23.0	2.1	1.8	---	5.1	---	50.4	0.4
NYTC(EN)	New York Times Corpus	**40.6**	2.2	1.5	0.4	13.5	8.0	44.1	0.2
LIC-2019(CN)	search results of Baidu Search as well as Baidu Zhidao	---	3.0	---	2.1	2.5	1.3	---	---