What is Special about Patent Information Extraction?

Liang Chen (Speaker)

Shuo Xu, Weijiao Shang, Zheng Wang, Chao Wei, Haiyun Xu

August 1st, 2020

1st Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents at the ACM/IEEE JCDL2020

Content

1. **Motivation**
2. Three particularities in PIE
3. Conclusion

1. Motivation

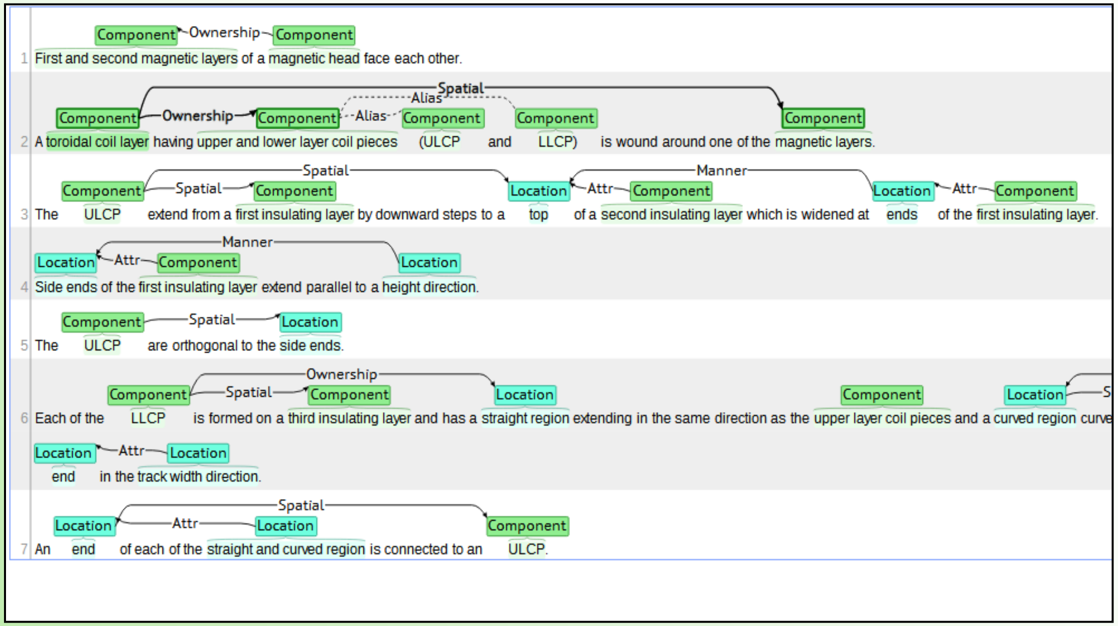
Why we discuss the specialty in patent information extraction?

About 3 years ago, with the support of NSFC (Natural Science Foundation of China), we began the research of patent information extraction

1. Motivation

Why we discuss the specialty in patent information extraction?

➢ built a patent labeled dataset pertaining to hard disk drive, namely TFH-2020 [1], which is provided for free download from *https://github.com/awesome-patent-mining/TFH_Annotated_Dataset*

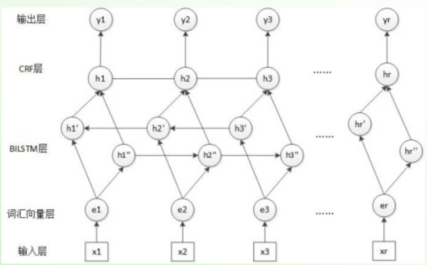


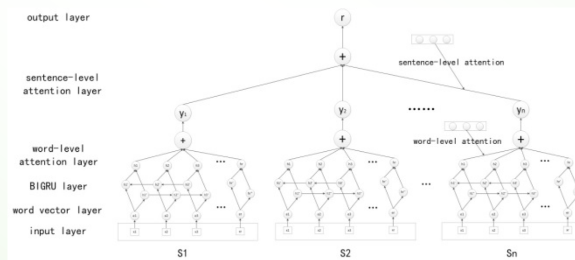
[1] Chen, L., Xu, S., Zhu, L. et al. A deep learning based method for extracting semantic information from patent documents. Scientometrics (2020). https://doi.org/10.1007/s11192-020-03634-y

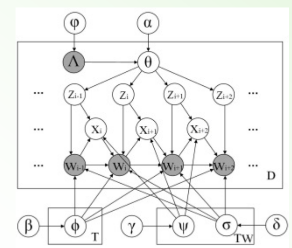
1. Motivation

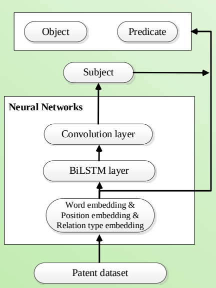
Why we discuss the specialty in patent information extraction?

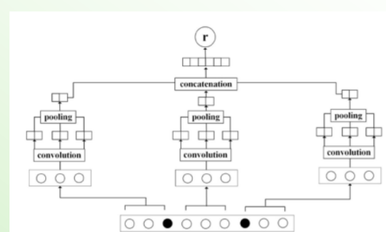
- ➢ employed a series of probabilistic graph models and deep learning models for patent information extraction;











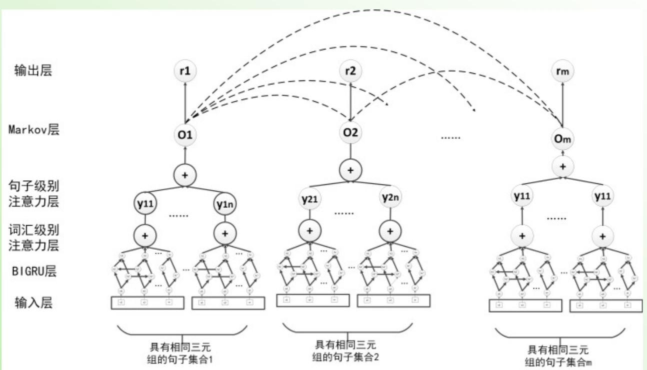
……

1. Motivation

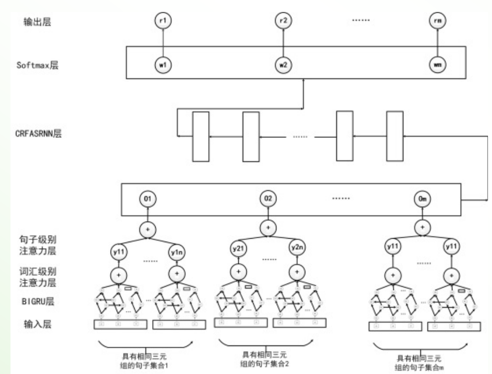
Why we discuss the specialty in patent information extraction?

- proposed several improved models with the specialty of patent text in concern.

Failed model 1



Failed model 2



Successful models

The paper is being prepared

1. Motivation

Why we discuss the specialty in patent information extraction?

- There are great differences between information extraction from patent text and generic text.
- Understanding these differences will effectively improve the performance of patent information extraction.
- Patent information extraction is a very big topic, so we only choose three aspects to share as follows.

Contents

2. Three particularities in Patent Information Extraction

2.1 The particularity of labeled patent dataset

At present, researchers' understanding of the specialty of patent text is mainly based on subjective judgment and simple investigations, like:

- Patent text follow a specific writing style[2]；
- The sentences in patent documents are more lengthy and syntactically complicated[3];
- ……

[2] Risch, J., & Krestel, R. (2019). Domain-specific word embeddings for patent classification. Data Technologies and Applications, 53(1), 108–122.

[3] Rajshekhar, K., Shalaby, W., & Zadrozny, W. (2016). Analytics in post-grant patent review: possibilities and challenges (preliminary report). In Proceedings of the American Society for Engineering Management 2016 international annual conference.

2. Three particularities in Patent Information Extraction

2.1 The particularity of labeled patent dataset

If we want to improve the models for patent information extraction, the specialty of patent text has to be cleared with data support.

To this end, 7 datasets are collected from different domains.

CPC-2014(EN)	CGP-2017(EN)	TFH-2020(EN)	Conll-2003(EN)	Wikigold(EN)	NYTC(EN)	LIC-2019(ZH)
Patent full-text regarding biology and chemistry	Patent abstract regarding biomed-	Patent abstract regarding thin film head techniques	Reuters news stories	Wikipedia	New York Times Corpus	search results of Baidu Search as well as Baidu Zhidao

Furthermore, 8 indicators are employed to analyze the datasets

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_i indicates 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_i indicates 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

2. Three particularities in Patent Information Extraction

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	0.3
TFH-2020(EN)	Patent abstract regarding thin film head techniques	30.7	**6.1**	**2.3**	**4.3**	2.8	1.2	**75.5**	**7.6**
Conll-2003(EN)	Reuters news stories	14.6	1.7	1.5	---	33.3	---	37.6	0.06
Wikigold(EN)	Wikipedia	23.0	2.1	1.8	---	5.1	---	50.4	0.6
NYTC(EN)	New York Times Corpus	**40.6**	2.2	1.5	0.4	13.5	8.0	44.1	0.04
LIC-2019(CN)	search results of Baidu Search as well as Baidu Zhidao	---	3.0	---	2.1	2.5	1.3	---	---

2. Three particularities in Patent Information Extraction

2.1 The particularity of labeled patent dataset

Conclusion:

- There exists difference between patent text and generic text;
- There exists significant difference between patent text from different technical domains;
- Such differences enable the performance of information extraction model to improve greatly.

2. Three particularities in Patent Information Extraction

2.1 The particularity of labeled patent dataset

- Due to the specialty of TFH-2020, our new model improved the relation classification by 3.2% in terms of micro-average F1-value, which is a remarkable improvement.

	micro-average(%)			macro-average(%)			weighted-average(%)		
	pre	rec	F1	pre	rec	F1	pre	rec	F1
WGCN	46.0	46.0	46.0	19.1	18.0	17.5	39.4	46.0	41.0
WLGCN	45.4	45.4	45.4	22.4	18.1	17.9	39.4	45.4	39.6
BiGRU-HAN	63.4	63.4	63.4	42.0	40.5	41.0	63.0	63.4	63.2
BiGRU-HAN-WGCN	**66.7**	**66.7**	**66.7**	**45.8**	**43.5**	**44.3**	**66.3**	**66.7**	**66.4**
BiGRU-HAN-WLGCN	66.1	66.1	66.1	45.3	43.0	44.0	65.6	66.1	65.8

2. Three particularities in Patent Information Extraction

2.2 The particularity of patent word embedding

When deep learning techniques are used for patent information extraction, a preliminary question is:

which kind of word embeddings should be used?

Notice! general speaking, patent information extraction is a domain-specific task

- Use word embedding trained on generic text?
- Use word embedding trained on patent texts from all fields?
- Use word embedding trained on patent texts from the same technical field?

WORD EMBEDDING

GloVe: Trained on generic texts

USPTO-5M: trained with the full-text of 5.4 million patents

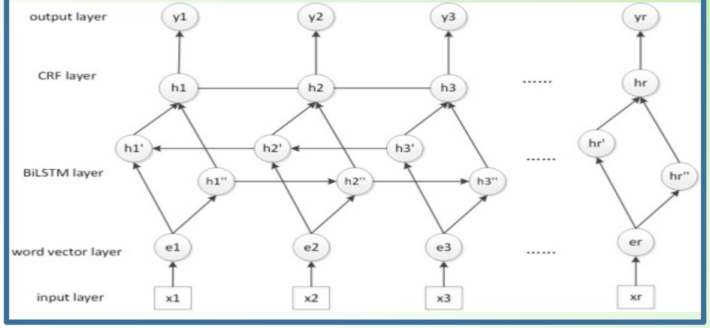
TFH-1010: trained on the full-text of 1010 patents from TFH datasets

TFH-46K: trained on the abstracts of 46,302 patents regarding magnetic head in hard disk drive

MODEL

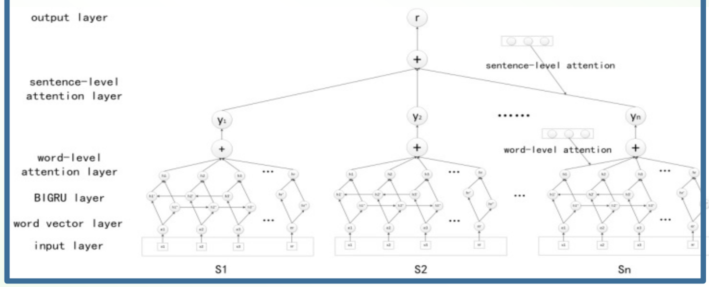
Named Entity Recognition

BiLSTM-CRF



Relation Extraction

BiGRU-HAN



2. Three particularities in Patent Information Extraction

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

2. Three particularities in Patent Information Extraction

2.2 The particularity of patent word embedding

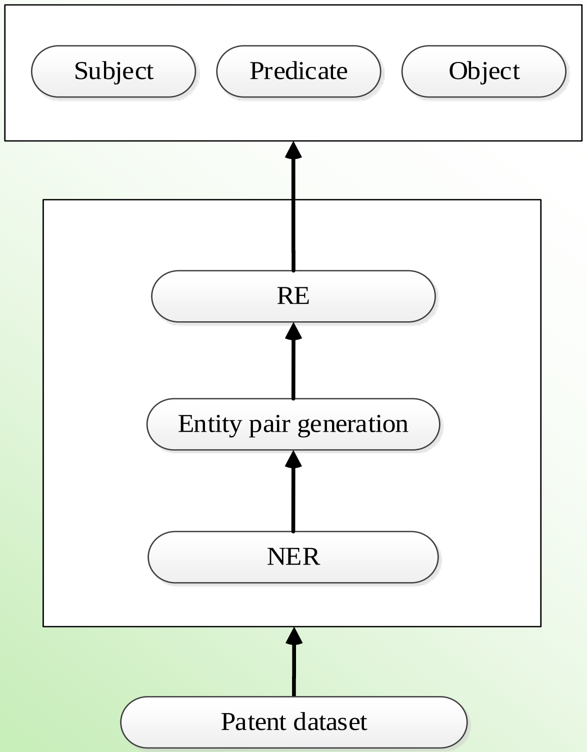
Conclusion:

- When extract information from patent text in certain domain, the word embedding trained on corpus from the same domain is preferable;
- If the scale of such corpus is limited, the addition of texts from relevant domain will help.

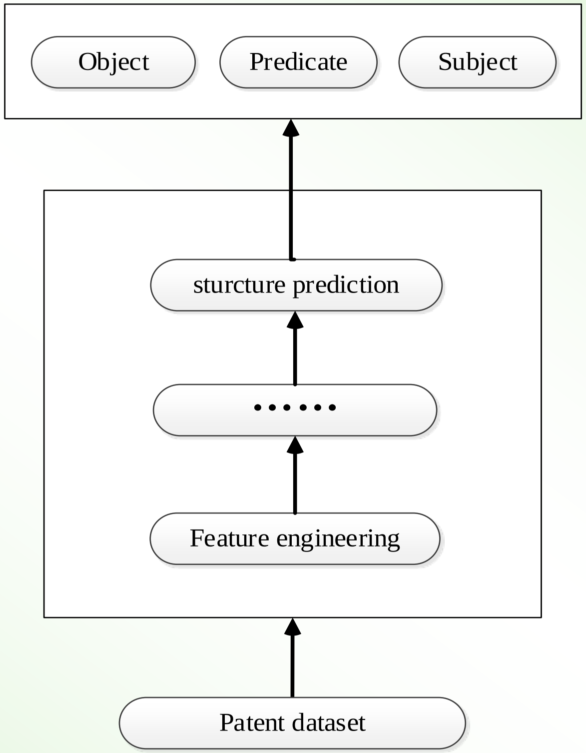
2. Three particularities in Patent Information Extraction

2.3 The particularity of method in patent information extraction

As state-of-the-art method in information extraction, sentence-level supervised learning method has 2 sub-classes, namely pipeline method and joint method.



(a) pipeline method



(b) joint method

2. Three particularities in Patent Information Extraction

lic2019.ccf.org.cn/kg

2019语言与智能技术竞赛 登录

比赛介绍 数据下载 结果提交 获奖名单&排行榜 新闻中心

Rank	Model	Precision	Recall	F1	Submit Time
1	[知识工场] BERT(ensemble) gdm 复旦大学	0.8975	0.8886	0.893	2019/5/20
2	[variant bert+multi head selection] (ensemble) littlebert 个人	0.8962	0.8886	0.8924	2019/5/20
3	[ERNIE CTagging + MultiSub Reviewer] (ensemble) Kill_Thread Ecole X	0.8976	0.8852	0.8914	2019/5/20
4	good luck(ensemble) 格物致知 国双科技	0.8948	0.8858	0.8903	2019/5/20

2. Three particularities in Patent Information Extraction

Aside from powerful models, the excellent performance also comes at the expense of large labeled dataset, which is far beyond the scale of labeled patent datasets available at present.

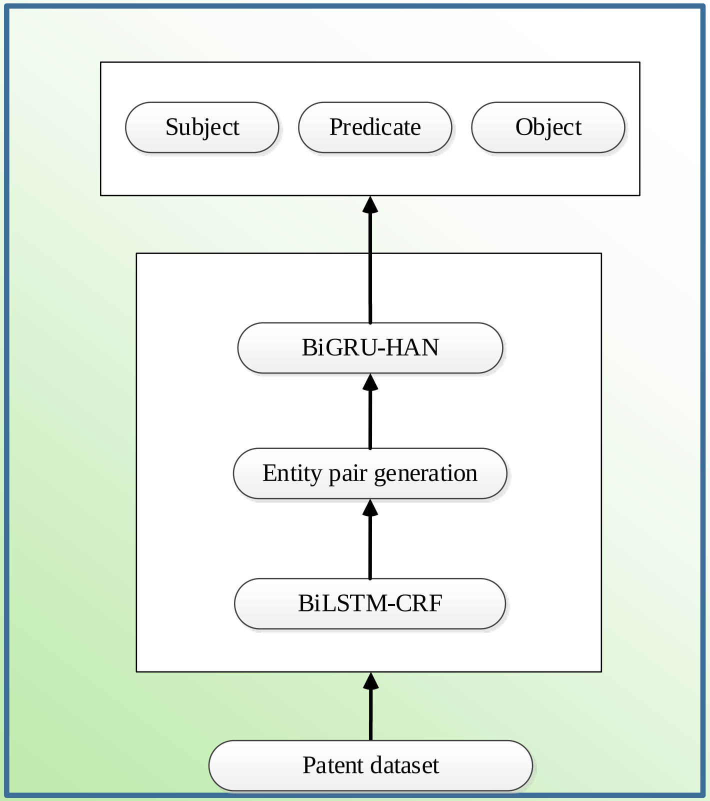
dataset	LIC-2019	CGP-2017	TFH-2020
# of instances	210,000	15,739	17,468

So how about the performance of different methods in patent information extraction?

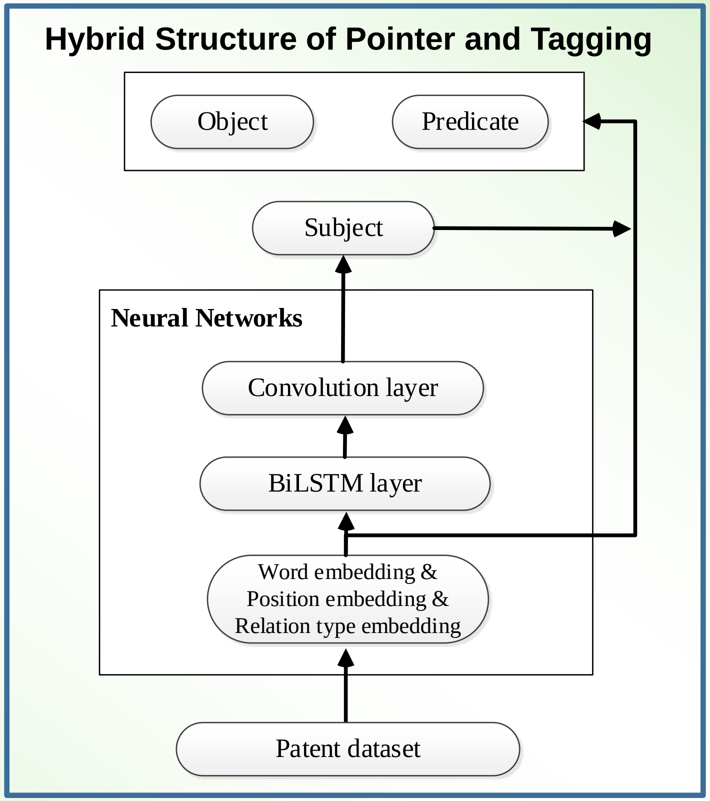
2. Three particularities in Patent Information Extraction

TFH-2020

Pipeline method



Joint method



2. Three particularities in Patent Information Extraction

The experimental results

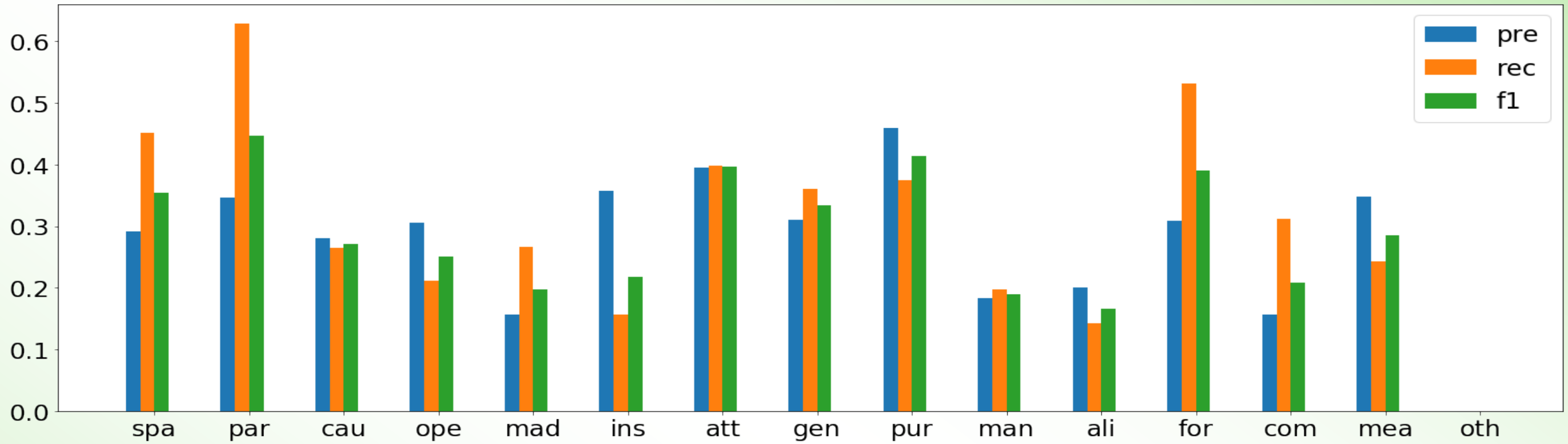


Fig.7 Result of **pipeline method** for information extraction

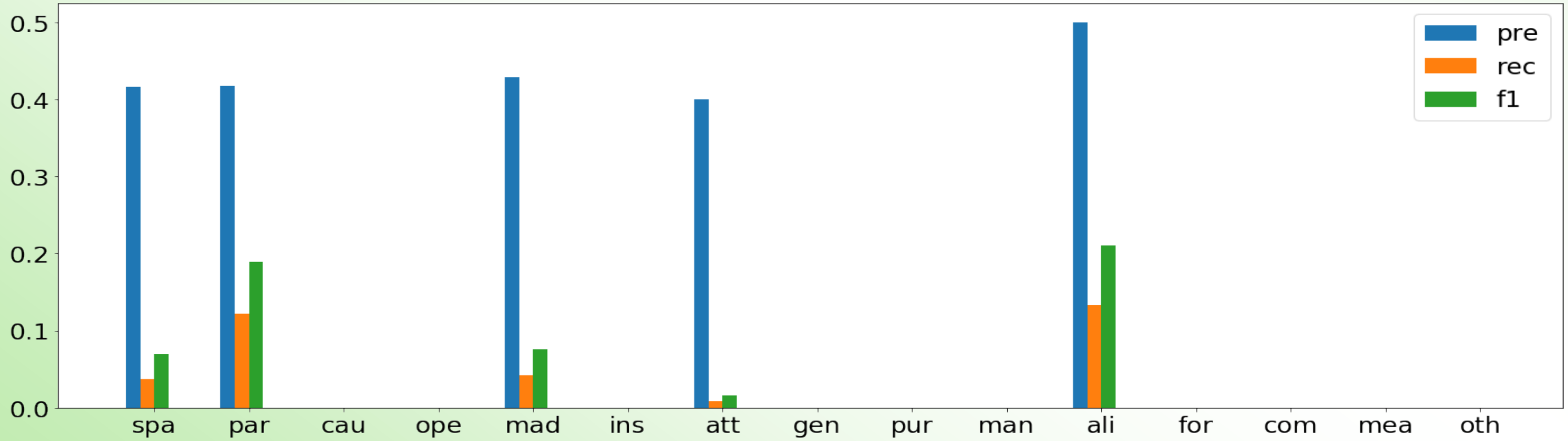


Fig.8 Result of **joint method** for information extraction

2. Three particularities in Patent Information Extraction

2.3 The particularity of method in patent information extraction

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) Furthermore, the performance of joint method is severely affected by the size of training set size.

2. Three particularities in Patent Information Extraction

How the size of training set affect the performance of *HSPT(a joint model*

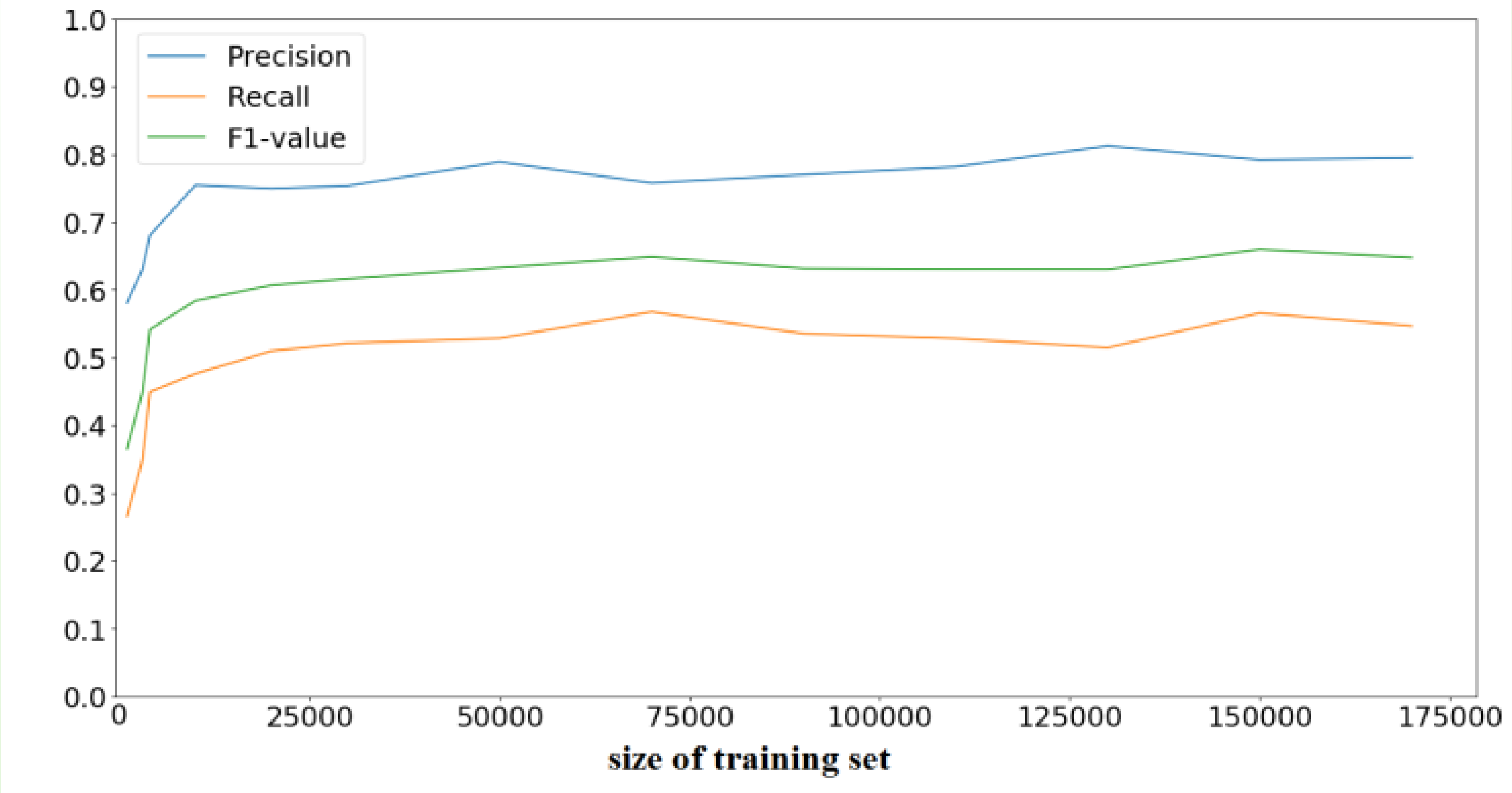


Fig.9 The performance of joint model with different size of training set

3. Conclusion

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

(1) Labeled dataset;
(2) Word embedding;
(3) Organization of sub-tasks in information extraction

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 publicly available so far, this is what we can get with data support.

Thanks!
Q&A

Email: 25565853@qq.com