Scientometrics (2020) 125:289–312
https://doi.org/10.1007/s11192-020-03634-y

A deep learning based method for extracting semantic information from patent documents

Liang Chen[1] · Shuo Xu[2] · Lijun Zhu[1] · Jing Zhang[1] · Xiaoping Lei[1] · Guancan Yang[3]

Received: 2 January 2020 / Published online: 24 July 2020

Abstract
The text-based patent analysis is grounded in information extraction technique. However, such technique suffers from obvious defects such as low degree of automation and unsatisfactory extraction accuracy. To deal with these problems, after an information schema is pre-defined, which contains 17 types of entities and 15 types of semantic relations, a dataset of 1010 patent abstracts is annotated and opened freely to the research community. Then, a novel patent information extraction framework is proposed, in which two deep-learning models, BiLSTM-CRF and BiGRU-HAN, are respectively used for entity identification and semantic relation extraction. Finally, to demonstrate the advantages of the new framework, extensive experiments are conducted, and the SAO method and PCNNs model are taken as respective baselines on the framework and module levels. Experimental results show that our framework out-performs the traditional one in terms of automation and accuracy, and is capable of extracting fine-grained structured information from patent texts.

Keywords Patent analysis · Entity identification · Relation extraction · Deep learning · BiGRU-HAN · BiLSTM-CRF · Thin film head · SAO · PCNNs

✉ Shuo Xu
xushuo@bjut.edu.cn

Liang Chen
25565853@qq.com

Lijun Zhu
zhulj@istic.ac.cn

Jing Zhang
janezhang@istic.ac.cn

Xiaoping Lei
leixp@istic.ac.cn

Guancan Yang
yanggc@ruc.edu.cn

[1] Institute of Scientific and Technical Information of China, Beijing 100038, People's Republic of China

[2] Research Base of Beijing Modern Manufacturing Development, College of Economics and Management, Beijing University of Technology, Beijing 100124, People's Republic of China

[3] School of Information Resource Management, Renmin University of China, Beijing 100872, People's Republic of China

Introduction

Patent document is a type of important intellectual resource, from which valuable technical intelligence can be obtained for technology opportunity discovery (Lee and Lee 2019), invention protection (Park et al. 2012), technology trend analysis (Han et al. 2017) and so on. As a matter of fact, so far technical intelligence is mainly obtained by expert reading (Yang 2012; Zhang 2016), which is laborious and inefficient, especially when the number of patents has been increasing dramatically due to rapid development in various technology areas in recent years. The automatic reading comprehension (Chen 2018) on the patents for technical intelligence becomes a significant challenge for the entire patent system.

Information extraction, armed with some powerful machine learning method, is one of the fundamental building blocks for computers to understand natural language, since it is capable of solving the ambiguous problem inherent in free texts by converting texts into semantic network (Singh 2018). However, just as Lupu (2017) noted, though such AI techniques have been widely applied to analyze large amounts of news (Phan and Sun 2018), scientific publications (Xu et al. 2015), electronic medical record (Ford et al. 2016) and so on, there are few successful cases from the IP (Intellectual Property) field in the literature. In our opinion, the main reasons are two-fold: (1) 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 (Rajshekhar et al. 2016); (2) Many information-extraction tools embed a supervised machine learning model, which relies on the availability of large annotated corpus from the target category of text resource, but the labeled patent dataset is public unavailable for a long time, especially from the fields other than biology.

A naïve way is to extract structured information from patent documents with a general-purpose tool, such as *Stanford CoreNLP* (Manning et al. 2014), *OpenNLP* (Baldridge 2005) and so on, and then to post-process the results with human-curated rules. However, the performance is not satisfactory in most cases (Souili et al. 2015; Rajshekhar et al. 2016; Carvalho et al. 2014) since the semantic information in patent documents don't usually follow the built-in patterns in most general-purpose tools. Especially when it comes to fine-grained information extraction with the focus on domain-specific entities and semantic relations, the general-purpose tools are even helpless. Furthermore, due to the lack of patent annotation dataset, a significant step, the quantitative evaluation of resulting proposed method, often misses from most previous research on patent information extraction. To alleviate the problem, a corpus from *thin film head* subfield in the field of *hard disk drive* is annotated comprehensively in this study and can be accessed public freely.[1] To the best of our knowledge, only two chemical patent corpora (Pérez-Pérez et al. 2017; Akhondi et al. 2014) are available in the literature, but only chemical and biological entities are annotated. In our corpus, in addition to the entity annotations, semantic relations between entities are annotated as well.

In addition, to benefit from the cutting edge NLP (Natural Language Processing) techniques for patent information extraction, a novel patent information extraction framework is proposed with leading deep learning models as its core modules to fulfill a series of sub-tasks. This is another contribution. Compared to previous methods, our framework has the following features: (1) The entities and semantic relations can be

[1] https://github.com/awesome-patent-mining/TFH_Annotated_Dataset.

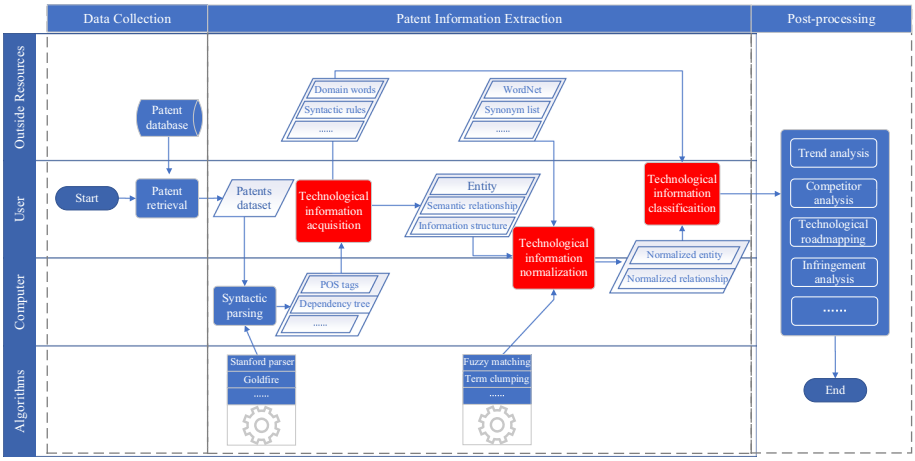

Fig. 1 Traditional patent information extraction framework

extracted simultaneously from patent documents effectively and efficiently; (2) Much more types of entities and semantic relations are supported. In fact, one can utilize user-definable entity and semantic relation types, as long as the fed dataset contains enough annotation information for model training; (3) The performance of patent information extraction method can be evaluated in a more objective and credible way.

The organization of the rest of this paper is as follows. After related work is briefly reviewed in Section *Related Work*, we create an annotated patent dataset for patent information extraction in Section *Data annotation*. Then, a novel framework of information extraction for patent documents is put forward in Section *Framework for Patent Information Extraction* and core modules are also described in more details in this section. Section *Experimental Results and Discussions* describes the detailed information of experiment, in which the SAO (Subject-Action-Object) method and PCNNs model are taken as baselines to demonstrate the advantages of our methodology. The last section concludes this contribution with future study directions.

Related work

Patent information extraction aims to solve a basic problem that has long plagued patent analysis, namely, how to efficiently and accurately identify terminologies from massive patent documents which cover technologies and their attributes, functions, effects, even corresponding products' name, thus to support further applications. Ever since pioneer work by Tsourikov et al. (2000), a variety of methods have been proposed, such as SAO method (Park et al. 2012), property-function method (Yoon and Kim 2012), and ontology-based method (Dewulf 2011). Currently the methods of patent information extraction are mainly oriented to practical applications and most of them are presented as a series of processing steps. The framework in Fig. 1 summarizes these steps, and the key modules, as shown in red box, will be described in the following subsections.

Technological information acquisition

Due to unavailability of information-extraction tools specifically oriented to patent documents, the goal of this step is to obtain structured information preliminarily with a general-purpose tool. Since text-mining tools can be easily accessed nowadays, they are usually used by IP researchers to generate PoS (Part-of-Speech) and syntactic dependencies of patent texts, and then certain terminologies and their combinations are filtered with a rules-matching method. It is not difficult to see that the key point of this process is how to curate manually these rules.

Let's take the SAO method as an example. Yang et al. (2017) were interested in the SAO structures consisting of Subjects (noun phrase), Actions (verb phrase), and Objects (noun phrase), since they argued that a SAO structure explicitly depicts a relation between components, in which the subject represents a type of technology, and the action and object are collectively viewed as a concept of function. Thus, a SAO structure should represent a problem that needs to be solved in the resulting patent documents (Moehrle et al. 2005). Along this direction, Wang et al. (2015) and Guo et al. (2016) used *Stanford Parser* (Chen 2018) to extract SAO structures with a set of customized rules. With the assistance of *GoldFire* (former name Knowledgist2.5™) (Invention Machine Corporation 2001), Choi et al. (2012, Choi et al. 2013 derived SAO structures and expressed them in tree and roadmap manner for technology planning. Park et al. (2013) focused on the functional information reflected by Action-Object (AO) combination in order to locate the key technologies in different industrial fields by calculating the distribution of AOs among these industrial fields.

Though the SAO method is widely used for patent information extraction, much valuable information is lost. For instance, the subject and object do not have explicit entity types, and the action can't indicate specific semantic relation between subject and object. To reduce information loss, Dewulf (2011) summarized a set of manual rules, e.g., function concepts mainly expressed in verbs and property concepts expressed in adjectives, to identify entities with interested types from patent documents. Yoon et al. (2012) further refined the PoS compositions of functions and attributes as "adjectives + nouns" and "verbs + nouns", and five Stanford typed dependencies that can be grammatically transformed into these interested PoS compositions were utilized to identify the entities. To recognize relation between entities, An et al. (2018) defined five semantic relations (*inclusion*, *objective*, *effect*, *process*, and *likeliness*), and used prepositions between these entities to determine their relations.

Technological information normalization

The obtained technological information structures from last subsection are not still ready for further patent analysis, as there are plenty of variants with the same or similar meaning among them. To deal with this problem, public available or self-built vocabulary or knowledge base is widely used to normalize these structures. More specifically, with the hierarchical and relational structure contained in the vocabulary, users can (1) easily judge whether two entities are synonyms or not by computing their semantic similarity (Xu et al. 2009), or by transforming each entity to its upper-level one in a concept hierarchy to check if they are matched (i.e., concept generalization) (Yoon et al. 2015),

and then (2) further infer whether two technical information structures have the same meaning (Wang et al. 2019).

For the analysis of the biotechnological patents, Bergmann et al. (2008) developed a domain-specific vocabulary based filter to modify SAO structures into more general ones. The WordNet (Miller 1995), a large knowledge base of English, is utilized by Yoon et al. (2015) to collapse AOs semantically identical into common AOs by concept generalization, and then these common AOs are seen as a fundamental mechanism for searching similar functions across different domains. As for strongly domain-specific abbreviations not included in WordNet, Choi et al. (2012) took such abbreviations as sets of synonyms, namely *synset*s in WordNet, and integrated them into WordNet.

In real-world applications, normalization of technological information structures using such methods is quite laborious and time-consuming, as the lexical resources have to be updated frequently to include the unknown entities from technology frontier. To alleviate the burden of the construction of lexical resource, Yang et al. (2018) used a fuzzy matching algorithm to fold SAO components with similar meaning. Choi et al. (2012) integrated a clustering algorithm with vocabulary construction. In more details, the WordNet-based measurement is used to calculate the similarities between SAO structures and the AOs and word phrases are further clustered into several groups for synonym folding.

Technological information classification

In order to highlight the patentability of an invention and ensure its stability against possible infringement or invalidity lawsuit in future, patent applicants must disclose as many technical details as possible when drafting their patent documents. This results in the technological information distributing at different levels of granularity even after normalization, which enables such information to be still hard for analysis and interpretation. To solve this problem, researchers generally categorize the obtained technological information into pre-defined classes to improve its interpretability.

Choi et al. (2012, 2013) took subjects, objects from SAO structures as entities and divided them into four types: *product*, *technology*, *material* and *technology attribute*. In the meanwhile, the action-object pairs were viewed as binary relations representing functions of inventions (such as purpose type, effect type and partative type function) in Choi et al. (2012, 2013). In the end, such entities and functions were assigned to appropriate types according to a set of self-built mapping rules (Choi et al. 2012, 2013). Likewise, three kinds of action-object pairs were defined in Yoon and Kim (2012): (1) object function representing products' core functions, (2) attribute function representing sub-functions, and (3) structural relation representing structural components. In their follow-up work (Yoon et al. 2015), each type of action-object pair was attached with certain PoS tags. In this way, with simple PoS tags matching they identified product's core function and supporting technologies for further analysis (Yoon et al. 2015). It is worth mentioning that the technological information classification schema varies by the fields. For example, six categories of SAO structures were defined for the domain of dye-sensitized solar cells, and each category of structures was attached with a high frequent word list in Wang et al. (2015).

General remarks

From the above review on related work, one can see that a lot of manual intervention makes patent information costly to obtain in practice. To say it in another way, patent information extraction is far from mature. To summarize, three limitations can be observed as follows.

Insufficient types of technological information

It is well known that there is extraordinary rich information contained in patent documents. However, the technological information types defined in the literature can seldom completely cover them. For example, the entities and semantic relations in the field of hard disk drive in computer can be categorized to 17 types (cf. Table 1) and 15 types (cf. Table 2), respectively. These figures are much larger than counterparts in the literature, which will no doubt loss seriously some valuable information.

Low degree of automation for patent information extraction

Too much manual intervention is involved during the procedure of patent information extraction, such as creating manual rules and updating synonym vocabulary. Compared to other applications of cutting edge NLP techniques in patent domain, such as patent classification (Li et al. 2018), patent landscaping (Choi et al. 2019), technology forecasting (Zhou et al. 2020) and word embedding training for patent documents (Risch and Krestel 2019a highly automatic framework with advanced techniques has yet to emerge in patent information extraction.

Unsatisfactory accuracy in patent information extraction

As a critical component, to our knowledge, neither open-source or commercial advanced text-mining model is catered specifically to take the characteristics of patent documents (Rajshekhar et al. 2016; Strzalkowski. 1999) into consideration until now. The scarcity of high-quality annotated patent dataset further aggravates the problem. These are the main reasons why the accuracy in patent information extraction is still unsatisfactory. Some researchers (e.g., Xu et al. 2019) are also aware of the problem, and try to use deep neural network for a better performance in patent information extraction. But only limited semantic information is extracted, such as 5 types of entities in Xu et al. (2019). So this study want to explore the potential of deep neural networks in extracting valuable information from patents once such models get supported by a high-quality annotated dataset.

Data annotation

To enrich the annotated patent corpora, the patents pertaining to *thin film head* technology in hard-disk are collected from the United States Patent and Trademark Office (USPTO) database. Our search strategy combines keywords, application time and patent citations. At the first stage, 137 seed patents are retrieved with the search statement of "ABST/'thin film head' AND APD/1/1/1976->31/12/2003". Through forward and backward citation to

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 value** and K value

these seed patents at the next stage, the patents dataset is extended to 2,048. After removing irrelevant patents, 1010 patents are kept as our dataset for future annotation.

Technological information definition

After extensive literature review (Yang and Soo 2012; Choi et al. 2012), expert consultation and understanding of the patent dataset, 12 types of entities and 11 types of semantic relations are defined here, as shown respectively in Table 1 and Table 2. At length, for purpose of describing the structure of an invention, the following entity types are usually involved: *system*, *component*, *function*, *effect*, *consequence*, *attribute*, *measurement*, *value*, *location*, *material*, *shape*, *scientific concept*. As for its working mechanism, one can usually delineate an invention with the following semantic relation types: *spatial relation*, *part-of*, *operation*, *generating*, *in-manner-of*, *made-of*, *comparison*, *measurement*, *causative relation*, *formation*, *purpose*. As a further supplement, another 4 types of entities (*physical flow*, *information flow*, *energy flow*, and *state*) and 2 types of semantic relations (*attribution* and *instance-of*) are defined to describe the characteristic and catalogue information of a technology. In addition, alias relation is utilized to connect an entity and its synonyms. To prevent exception instances (viz., instances uncovered by above information types) of the entities and semantic relations, *other* type is appended to Tables 1 and 2.

Such a refined technological information classification schema can help users to assign a clear and professional tag to an interested entity and semantic relation, thus to improve the performance of patent analysis. Nevertheless, it is worth noting that it is not trivial to obtain a suitable technological information classification schema. As a thumb of rule, 3–4 iterations of a trial and error process are usually needed. That is to say, once a schema draft is done, the applicability testing alternatives the schema revision until a final version of schema.

Data labeling

Our team consisting of 10 members spends almost 6 months on the data labeling task. The biggest challenge in this process is how to ensure the consistency of 10 independent annotators. Here, an online-offline working mode is adopted. By online, we mean that a web-based annotation tool *Brat* (Stenetorp et al. 2012) is deployed on a cloud server in advance, and each annotator is allocated an account. The first author is responsible of allocating patent subset to each annotator and reviewing the annotated patents by each annotator. The qualified labeled patents are stored directly in tagged patent pool and unqualified ones are sent back to resulting annotator for re-labeling.

Another measure to ensure the consistency is the offline annotation process management. Firstly, all annotators are trained and assessed before formally labeling patents. Those who pass the assessment will receive 10–20 patent abstracts to tag weekly. Secondly, the first author reviews the annotation results by sampling and holds a discussion meeting weekly to comment and correct typical errors. Thirdly, each annotator is ranked based on his/her annotation results monthly. The workload for lower ranked annotators will be reduced accordingly.

In the end, 22,742 entity mentions and 17,421 semantic relation mentions are obtained in total. Our corpus includes 3,981 sentences with an average sentence length of 23.2 words. The distributions of entity and semantic relation types are illustrated in Figs. 2 and 3, respectively. From Figs. 2 and 3, one can observe that this dataset is very skewed for

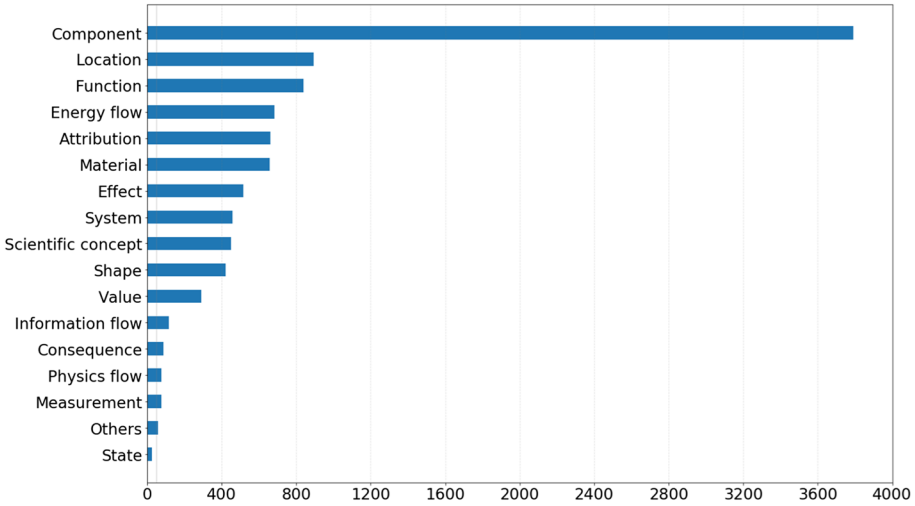

Fig. 2 The distribution of different entity types

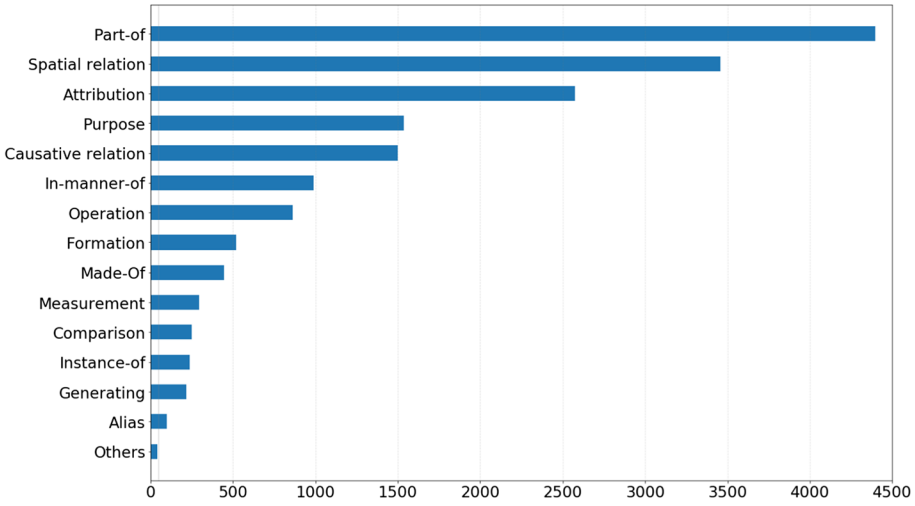

Fig. 3 The distribution of different semantic relation types

both entity and semantic relation types. As for entities, the most significant type is *component* which accounts for 37.5% of the total number, and then followed by *location*, *function energy flow*, and *attribution.* This conforms to our intuition that the main content in a patent is often about the description of an invention's structure, working mechanism, and the location, attribution, usage of its components. On the other side, it is well-documented that semantic relation types are closely related to the entity types. Hence, the *part-of* and *spatial* relations are top 2 largest ones, since these two relations devote to describe the connection between *component*s, and the *component* is the largest number of entity type amongst all entities. Similar phenomena are also observed for other sematic relation types

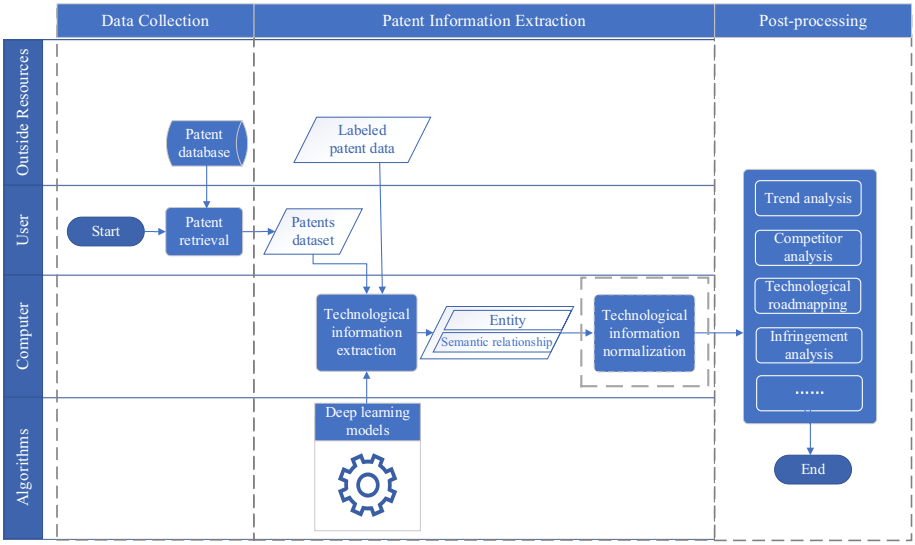

Fig. 4 A novel framework of patent information extraction

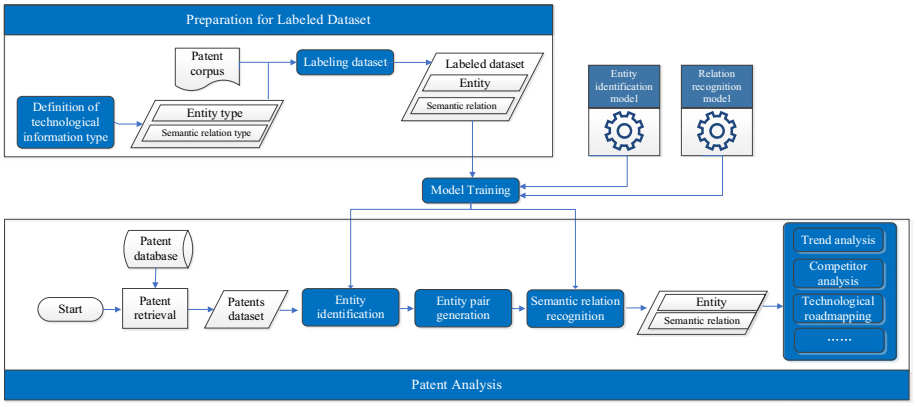

Fig. 5 The implementation of new framework

like *attribution*, *purpose*, and *causative* relation, since the involved entity types such as *attribution* and *function* also appear in high frequency.

Framework for patent information extraction

A new framework for patent information extraction is proposed, as shown in Fig. 4, in which the information extraction process is divided into a series of sub-tasks organized by a pipeline. Since the sub-tasks are clearly defined and functional independent to each other, the new framework achieves much better modularity than the previous one in Fig. 1. Moreover, this framework gets rid of manual rules, domain vocabulary, and human intervention.

Side	ends	of	the	insulating	layer
B-location	I-location	O	O	B-component	I-component

Fig. 6 An example for entity identification with BIO notation schema

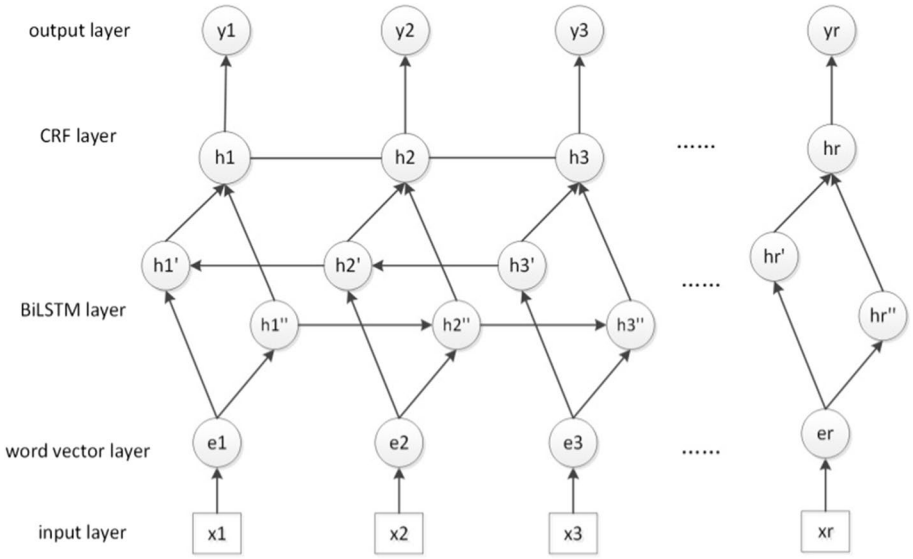

Fig. 7 The graph model representation of BiLSTM-CRF model

Herein only annotated dataset is needed for training an information extraction model, after which valuable patent information can be extracted automatically with high accuracy.

To verify the characteristics mentioned above, we provide a specific implementation of this framework as shown Fig. 5. After the types of technical information are defined and patent documents are annotated accordingly, a subset of the labeled dataset is randomly chosen for training an information extraction model. It is better for our framework to deploy a deep learning model. Last but not the least, with the help of the trained model, the structured information can be extracted from unseen patent documents. Note that the extracted information still needs to be normalized in our framework before further patent analysis. That is, this module is same as that in traditional framework in Fig. 1, so it is not shown in Fig. 5. In the following subsections, the key modules (entity and semantic relation recognition) of the process will be described one by one. For more elaborate and detailed description on how to define technological information types and annotate patent documents, we refer the readers to Section *Data Annotation*.

Entity identification

Entity identification, also known as named entity recognition (NER), seeks to locate and classify named entity mentions in unstructured text into pre-defined categories. In machine learning, entity identification is often seen as a sequence labeling problem, in which *B* is

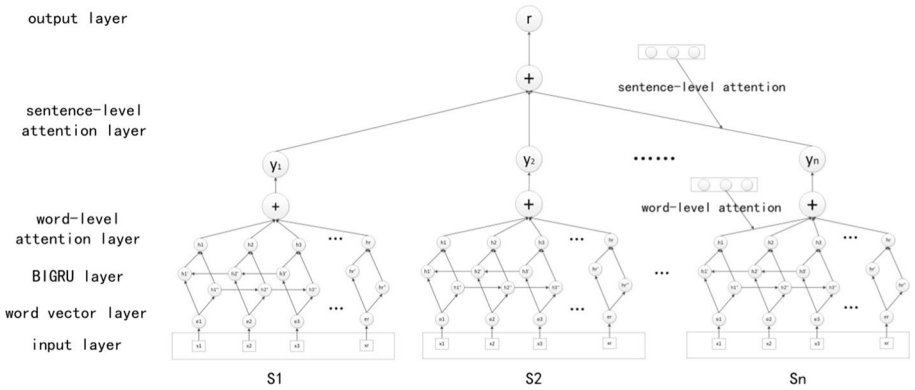

Fig. 8 The graph model representation of BiGRU-HAN model

attached to the tokens as the beginning of entities, *I* to those as the inside of entities and *O* for non-entity tokens. For convenience of understanding, an example from our corpus is shown in Fig. 6. In this case, there are two types (*component* and *location*) of entities. To distinguish entity types, type name is appended to the corresponding tags.

As one of the state-of-the-art deep neural network models for entity identification, BiLSTM-CRF (Huang and Xu 2015) shown in Fig. 7 is used in our framework. This model takes sentences as input and represents every word as a vector named word embedding, during training procedure these word embeddings pass through the layers within BiLSTM-CRF and output the predicted label for each word in the sentence. With the help of back propagation algorithm, the predicted labels will approximate the true labels and finally enable BiLSTM-CRF to recognize named entities in new sentences.

Entity pair generation

Before extracting semantic relations, binary-relations between entities need to be acquired. As we all know, a sentence may contain several entity mentions, but a specific semantic relation only holds for some entity pairs. In order to filter out entity pairs which are obviously impossible to form an interested semantic relation, the following selection preference (Jurafsky and Martin 2019) is used here. A set of rules are built ahead by enumerating all possible valid combinations of entity types. The entity pairs that do not meet any rule will be excluded directly from further analysis. For illustration, suppose we only have *made-of* relation type which requires an entity pair of (*component, material)*, then any entity pairs not matched will be removed from the entity pair samples. After filtering, it is very possible that some entity pairs are kept, but are not annotated by any semantic relation type. In this case, a special type, *no relation*, will be assigned to them. In semantic relation extraction step, the *no relation* is treated as a common label among other relation types for deep learning model to classify.

Semantic relation extraction

Similar to entity identification, another deep neural network model, BiGRU-HAN (Han et al. 2019), is utilized in our framework. This is a dedicated model for semantic relation

extraction, and it is made up of six layers as shown in Fig. 8. The basic idea of BiGRU-HAN 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.

Experimental results and discussions

Even though our annotation dataset is the first public available one with both entities and semantic relations labeled in patent domain, the volume of annotated patents is still limited, compared to other general-purpose annotation datasets, such as the English labeled dataset of CoNLL-2003 Shared Task with 22,137 sentences (Sang and De Meulder 2003) and the labeled dataset of 2019 Language and Intelligence Challenge with 21 thousand of sentences (Wu. 2019). Since the conventional way of splitting corpus into training set, validation set and test set will further reduce the size of training set, herein the corpus is splitted randomly into a training set and a test set with a 4:1 ratio in the patent level, which include 3259 and 722 sentences respectively. To tune hyperparameters in deep learning models, the dataset from CoNLL-2003 Shared Task is utilized and then transferred to the patent dataset, which is a common practice in deep learning model training.

In addition, a commonly used method of patent information extraction, namely SAO, is taken as baseline to illustrate the advantages of the new framework. To keep the comparison fair, all extracted information is kept intact. That is, technological information normalization module is removed from this comparison. In addition, to avoid the influence of subjective factors (such as manually-curated rules in the SAO method), the classification performance is not compared in this study. In this way, technical information types are not taken into consideration in the performance comparison for identifying entity and semantic relation.

Word embeddings

There are two ways to obtain word embeddings (1) by training on a corpus via word embedding algorithm, such as Skip-gram (Mikolov et al. 2013), CBOW (Mikolov et al. 2013) and the like; (2) by directly downloading a pre-trained word embedding file from the Internet, like GloVe (Pennington et al. 2014). Risch and Krestel (2019) 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 our information extraction from patents in specific domain. In order to explore which word embedding is preferable in our task, 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 (Pennington et al. 2014). 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;

Table 3 The summary of entity identification 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

2. Word embeddings provided by Risch and Krestel (2019), 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-5 M;
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-46 K.

On the basis of these word embeddings, we ran our methodology, but the results produced by these four types of word embedding are almost the same. Due to space limitation, only the performance for the entity identification is shown in Table 3.

However, as Risch and Krestel (2019) 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.

Therefore, we mainly describe this work using word embeddings from training on the abstract of 46,302 patents regarding magnetic head in hard disk drive instead. Specifically, we used CBOW algorithm implementation in Gensim tookit[2] and set the window size/minimum word frequency to be10/5 respectively. The number of epochs is fixed to 5. This setting is the most used configuration in word embedding training.

Entity identification

After 20 epochs of model training, weighted-average precision, recall, F1-value of BiLSTM-CRF on entity-level for the test set are 78.5%, 78.0%, and 78.2%, respectively. Although such performance is acceptable, it is still lower than its performance on general-purpose dataset by more than 10% in F1-value. The main reason is the limited amount of labeled dataset. Figure 9 shows precision, recall,and F1-value for each type of entity

[2] https://radimrehurek.com/gensim/.

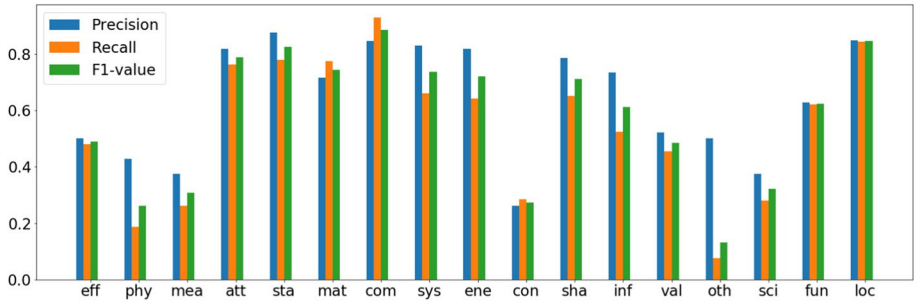

Fig. 9 Result of entity identification for different entity type

Table 4 The statistics for each type of entity

Type	#of correctly identified ones	Support	Type	#of correctly identified ones	Support
Effect	61	127	Consequence	6	21
Physical flow	6	32	Shape	95	146
Measurement	6	23	Information flow	22	42
Attribution	242	318	Value	25	55
State	7	9	Other	1	13
Material	211	273	Scientific concept	36	128
Component	1817	1958	Function	180	290
System	121	183	Location	312	370
Energy flow	176	274			

Gold Standard	The	**inductive head**	includes	a	**leading write pole**	and	a	**trailing write pole**
Extracted Result	The	**inductive head**	includes	a	leading **write pole**	and	**a**	**trailing** write pole

Fig. 10 An example sentence with gold standard and extracted results

denoted by its first 3 letters (cf. Table 1), and the confusion matrix is illustrated in Appendix. Furthermore, we display the number of each type of entities, namely support degree, in the test dataset and of the correctly identified ones in Table 4. From Fig. 9, the performance varies by entity type. For example, the best performance is for *component* type in term of F1-value (88.5%), while for *consequence* and *other* types, the F1-value dramatically fall down to below 30%. One interesting phenomenon can be observed that the entity identification performance in term of F1 score is positively correlated to support degree, such as 1,958 support degree for *component* type versus 21 and 13 support degree for *consequence* and *other* type. This indicates that one can improve the entity identification performance by promoting corresponding support degree.

As to SAO method, three strategies are adopted for result evaluation, (1) Exact match: a correct identification only happens when an extracted entity is exactly matched with gold standard in annotated corpus; (2) Inclusion match: an extracted entity is regarded to be correct if the entity includes a counterpart in labeled dataset or is included by some

Table 5 The summary of entity identification result

	Precision (%)	Recall (%)	F1-value (%)
SAO method			
Exact match	3.0	1.2	1.7
Inclusion match	14.8	5.7	8.3
Overlap match	62.3	28.2	38.8
Our framework			
Exact match	92.4	91.9	92.2

counterpart; (3) Overlap match: one can view an extracted entity as correct one when it overlaps with a counterpart from benchmark dataset. For illustration, the sentence "The inductive head includes a leading write pole and a trailing write pole" is taken as an example, as shown in Fig. 10. This sentence mentions three entities, *inductive head*, *leading write pole*, and *trailing write pole*. According to exact match strategy, only *inductive head* is correctly identified, but if one switches to inclusion match strategy, *write pole* is also regarded to be correct. Of course, if the strategy is relaxed further to overlap match strategy, all 3 entities are treated as correctly identified mentions.

Table 5 reports the results from our method and SAO method. Note that since SAO method only outputs entity boundary without resulting type. For convenience of comparison, only entity boundary information is used to evaluate BiLSTM-CRF model in Table 5. But even with the strictest strategy, namely extract match strategy, a huge gap between these two methods can also be observed from Table 4. For SAO method, 71.8% of the gold standard entities are not identified at all. In more details, only 1.2%, 4.5% and 22.5% of the gold standard entities exactly, inclusively and overlapping match the outputs of SAO method, respectively. From another perspective, among the output entities of the SAO method, totally incorrect entities accounts for 37.7%, in which exactly, inclusively and overlapping matched entities respectively make up 3.0%, 11.8% and 47.5% with the gold standard entities. In contrast, the performance of our framework is much better than that, as even in exact match way, it can identify 91.9% of the gold standard entities, and the correct entities account for 92.4% of the total entities it outputs.

Semantic relation recognition

According to a set of rules such as selectional preference rules, avoidance of the combination of entities and themselves, there are 157,372 entity pairs generated for semantic relation recognition, among which training set contains 129,976 entity pairs with 115,533 *no relations*, and test set 27,396 entity pairs with 24,418 *no relations.*

Overall evaluation

After 20 epochs of training, the average result of BiGRU-HAN on the test dataset is shown in the first row of Table 6. Note that these figures are obtained by considering simultaneously *no relation* accounting for 89.1% relation instances. If these relations are excluded, the average result will drop to the third row of Table 6, and it reflects the capability of BiGRU-HAN in patent relation recognition more truthfully.

Table 6 The overall evaluation for BiGRU-HAN and PCNNs

	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
PCNNs with *no relation*	89.0	89.0	89.0	10.9	6.4	6.2	81.1	0.89	84.0
PCNNs without *no relation*	1.4	1.4	1.4	5.7	0.2	0.3	26.8	0.6	1.1

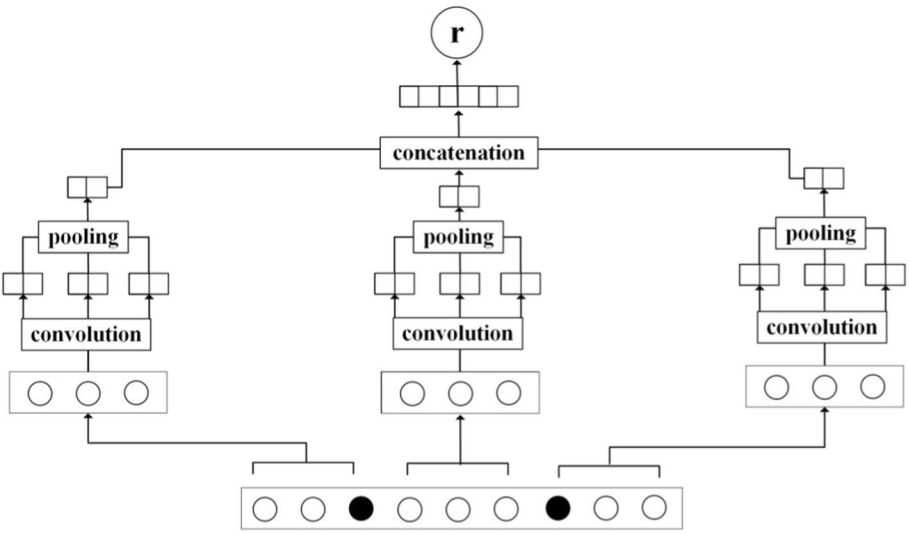

Fig. 11 The graph model representation of PCNNs model

In fact, not all state-of-the-art models can achieve such an performance in patent documents, the reason is that relation recognition for patents is quite a domain-specific task, as patent contains much more entities than generic text such as newspaper, Wikipedia, so after entity pair generation, the proportion of *no relation* is much larger than that of generic text, However, most state-of-the-art models can't be properly trained with such extremely imbalanced samples.

In order to illustrate this problem, we take another state-of-the-art model named PCNNs (Zeng et al. 2015) for example, the structure of this neural network is shown in Fig. 11 and its basic idea is to divide the input sentence into three segments by the two entities in an entity pair, and then uses the convolutional neural network to learn features from each segment and concatenate them into a larger vector for relation classification. After 20 epochs of training, the average results of PCNNs with and without *no relation* on the test dataset is shown in the second and forth row of Table 6, which

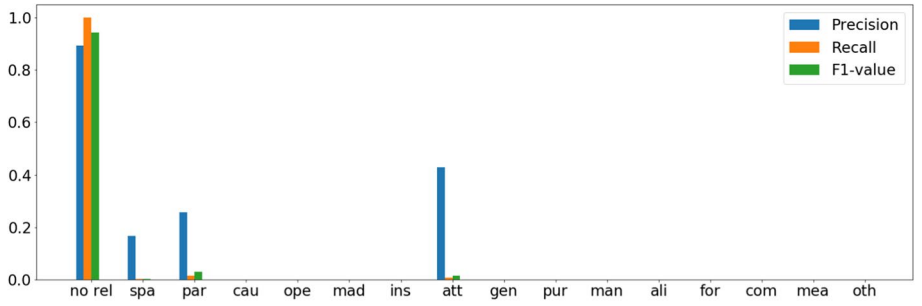

Fig. 12 Result of PCNNs for semantic relation recognition

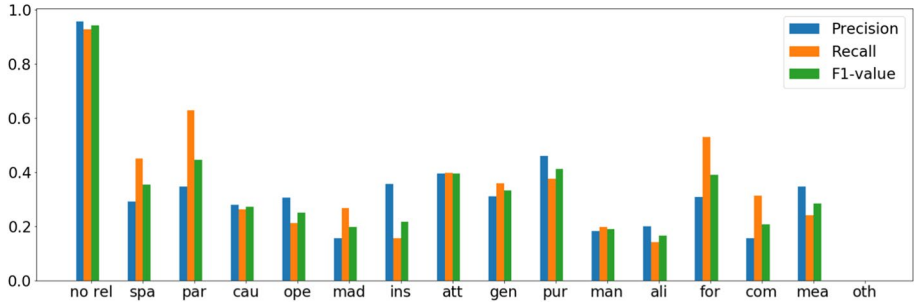

Fig. 13 Result of BiGRU-HAN for semantic relation recognition

Table 7 The statistics for each type of relation

Type	#of correctly identified ones	SUPPORT	Type	#of correctly identified onese	Support
No relation	12598	13580	Generating	9	25
Spatial relation	94	208	Purpose	57	152
Part-of	237	377	In-manner-of	15	76
Causative relation	47	178	Alias	2	14
Operation	15	71	Formation	17	32
Made-of	8	30	Comparison	5	16
Instance-of	5	32	Measurement	8	33
Attribution	66	166	Other	0	1

are much worse than that of BiGRU-HAN. On close examination, we find that PCNNs almost have no effect on other types of relation except *no relation* type, as shown in Fig. 12.

It's worth noting that even BiGRU-HAN is totally different from PCNNs, since they interact with outside environment in a consistent way (both take entity pairs with the sentences they belong to as input and take the predicted relation types as output). One can easily replace one model with another without modifying the other parts of framework, which as a result verifies the modularity of the new framework.

Detailed evaluation

In order to highlight the performance of BiGRU-HAN model on different types of relations, the precision, recall, and F1-value for each type of relation is shown in Fig. 13, and the counterparts' support degree and the number of correctly identified relations is shown in Table 7.

As we can see, the performance of relation recognition for each relation type is also positively correlated to its support degree. The *ownership* and *spatial* relation with top 2 support degrees achieve the best recognition performance, and for relation type with low support degree such as *generating* and *others* relation, their F1-values fall down to below 50%. It is worth noting that even with the similar support degree, the recognition performance for different relation types still varies considerably due to the various degree of difficulty for classifying different types of relation. Some relation types have obvious signals and can be easily determined, such as "of" for attribution relation, but *others* are too ambiguous even for humans to determine their type quickly.

Given that semantic relation recognition is based on the output of entity identification, the incorrect identification of entity will inevitably lead to the error of semantic relation recognition. As the entity identification performance of the SAO method is so poor under exact match and inclusion match strategies, it hardly supports the follow-up semantic relation recognition. Therefore, the overlap match strategy is adopted for the SAO method. That is, as long as the output entity overlaps a gold standard entity, the resulting entity is supposed to be correctly identified. Similarly, if two entities of a gold standard relation overlap those of a SAO structure, the relation is considered to be correctly identified.

With the above rule, the performance of semantic relation recognition for the SAO method is 41.6%, 13.4%, and 20.3% in terms of precision, recall, and F1-value, respectively. The performance of our framework is 45.8%, 58.8%, and 51.5%. Compared with entity identification, the performance gap of semantic relation recognition between SAO and deep learning method is slightly reduced, especially in term of precision. Our framework obviously out-performs the SAO method in term of recall. We argue that as a syntactic dependency-based information extraction method, the SAO method can only recognize a part of semantic relations that conform to ruled patterns, but it cannot effectively recognize other semantic relations beyond these patterns.

Conclusion

With the emergence of promising techniques, such as deep learning, information extraction for news, scientific publications, and electronic medical record has made great progress. However, information extraction for patent documents still suffers from low degree of automation, unsatisfactory accuracy and other deficiencies, which limit its further application. In order to deal with these problems, a novel framework is put forward in this paper after a patent corpus is annotated with the following contributions:

1. A set of patent documents pertaining to *thin film head* technology in hard-disk is labeled and shared in this work. To the best of our knowledge, this is the first labeled patent dataset in technology management domain that annotates both entities and the semantic relations between entities. Moreover, the well-crafted information schema used for patent annotation contains 17 types of entities and 15 types of semantic relations. This far exceeds the number of information types in the literature, thus to guarantee a strong support for subsequent patent analysis.
2. A novel patent information extraction framework with supervised learning and deep learning techniques as the core modules is raised in this study. Compared to the conventional methods, new framework is capable of learning the information extraction rules from labeled dataset effectively and efficiently. This means that manual interventions can be reduced dramatically. Thus, our framework can be competent in the task of information extraction from massive patent texts;

Though, there is still some room to improve our framework as follows. (1) As mentioned in Section *Experimental Results & Discussion*, the performance of entity identification in our framework is lower than on general-purpose information extraction counterparts. The main reason is insufficient labeled data, but dataset annotation is quite costly. Hence, how to generate automatically high-quality labeled dataset is a necessary work for patent information extraction. (2) Another problem in our framework lies in the imbalanced samples from the generation of *no* relation instances, which jeopardizes information extraction performance. At present, the entity pairs have not been generated in advance in the unified models (Zheng et al. 2017; Wu 2019). They can extract simultaneously the entities and their relations, and show a promising performance. In the near future, a unified model will be incorporated into our framework.

Acknowledgements 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.

Appendix

There are two types of errors for entity identification: (1) errors in entity boundary detection, (2) errors in entity type classification. General confusion matrix is capable of recording the first type of errors. As for the second type, an extra column is appended to the confusion matrix in Table 8, where rows indicate true entity types and columns predicted ones, and the last column (*ebd*) denotes the errors in boundary detection.

Table 8 The confusion matrix of entity identification

	eff	phy	mea	att	sta	mat	com	sys	ene	con	sha	inf	val	oth	sci	fun	loc	ebd
eff	61	0	0	6	0	0	4	0	7	6	2	1	0	0	5	11	0	24
phy	0	6	0	0	0	14	5	0	0	0	0	0	0	0	0	5	0	2
mea	1	0	6	5	0	0	0	0	0	0	0	0	3	1	2	1	0	4
att	3	0	9	242	0	1	15	0	1	1	2	0	0	0	13	3	11	17
sta	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	1	0	1
mat	1	2	0	1	0	221	18	2	0	0	0	0	0	0	0	0	0	38
com	2	2	1	5	0	5	1817	12	7	0	11	0	0	0	6	4	15	71
sys	0	0	0	1	0	0	44	121	0	0	1	0	0	0	0	2	0	14
ene	8	1	0	0	0	2	49	0	176	0	3	0	0	0	8	8	1	18
con	2	0	0	1	0	0	0	0	1	6	0	0	0	0	2	5	0	4
sha	2	0	0	3	0	1	24	1	0	0	95	0	0	0	2	2	4	12
inf	1	0	0	0	0	0	2	0	2	0	0	22	0	0	1	1	0	13
val	3	0	0	0	0	0	0	0	0	0	0	0	25	0	5	1	0	21
oth	0	0	0	0	0	1	3	0	0	0	0	0	0	1	0	1	0	7
sci	12	0	0	9	0	3	21	1	6	4	0	1	2	0	36	9	2	22
fun	7	2	0	2	1	1	27	0	4	3	0	0	0	0	7	180	0	56
loc	1	0	0	6	0	0	23	0	0	1	2	0	0	0	0	2	312	23

References

Akhondi, S. A., Klenner, A. G., Tyrchan, C., Manchala, A. K., Boppana, K., Lowe, D., et al. (2014). Annotated chemical patent corpus: A gold standard for text mining. *PLoS ONE, 9*(9), 1–8.

An, J., Kim, K., Mortara, L., & Lee, S. (2018). Deriving technology intelligence from patents: Preposition-based semantic analysis. *Journal of Informetrics, 12*(1), 217–236.

Baldridge, J. (2005). The OpenNLP project. http://opennlp.apache.org/index.html. Accessed 14 Dec 2019.

Bergmann, I., Butzke, D., Walter, L., Fuerste, J. P., & Erdmann, V. A. (2008). Evaluating the risk of patent infringement by means of semantic patent analysis: The case of DNA chips. *R& D Management, 38*(5).

Carvalho, D. S., França, F. M. G., & Lima, P. M. V. (2014). Extracting semantic information from patent claims using phrasal structure annotations. In *2014 Brazilian Conference on Intelligent Systems* (pp. 31–36).

Chen, D. (2018). *Neural reading comprehension and beyond (Doctoral dissertation)*. Palo Alto, CA: Stanford University.

Choi, S., Kang, D., Lim, J., & Kim, K. (2012a). A fact-oriented ontological approach to SAO-based function modeling of patents for implementing function-based technology database. *Expert System with Application, 39*(10), 9129–9140.

Choi, S., Kim, H., Yoon, J., Kim, K., & Lee, J. Y. (2013). An sao-based text-mining approach for technology roadmapping using patent information. *R&D management, 43*(1), 52–74.

Choi, S., Lee, H., Park, E. L., & Choi, S. (2019). Deep patent landscaping model using transformer and graph embedding. arXiv preprint arXiv: 1903.05823v4

Choi, S., Park, H., Kang, D., Lee, J. Y., & Kim, K. (2012b). An SAO-based text mining approach to building a technology tree for technology planning. *Expert Systems with Applications, 39*(13), 11443–11455.

Dewulf, S. (2011). Directed variation of properties for new or improved function product DNA- a base for connect and develop. *Procedia Engineering, 9,* 646–652.

Ford, E., Carroll, J. A., Smith, H. E., Scott, D., & Cassell, J. A. (2016). Extracting information fro-m the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association, 23*(5), 1007–1015.

Guo, J., Wang, X., Li, Q., & Zhu, D. (2016). Subject- action- object- based morphology analysis for determining the direction of technological change. *Technological Forecasting and Social Change, 105,* 27–40.

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

Han, C., Lim, H., Lee, D., Cho, H., & Kang, K. (2017). Patent analysis for forecasting promising technology in high-rise building construction. *Technological Forecasting and Social Change, 128*(3), 144–153.

Huang, Z., Xu, W., &Yu K. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.

Invention Machine Corporation. (2001). Knowledgist 2.5-Product Description http://www.triz.ch/KN25Prodesc.doc. Accessed 14 Dec 2019.

Jurafsky, D., Martin, J. (2019). Speech and language processing (the 3nd edition draft). https://web.stanford.edu/~jurafsky/slp3/. Accessed 24 Dec 2019.

Lee, C., & Lee, G. (2019). Technology opportunity analysis based on recombinant search patent landscape analysis for idea generation. *Scientometrics, 121*(2), 603–632.

Li, S., Hu, J., Cui, Y., & Hu, J. (2018). DeepPatent: patent classification with convolutional neural networks and word embedding. *Scientometrics, 117*(2), 721–744.

Lupu, M. (2017). Information retrieval, machine learning, and NLP for intellectual property information. *World Patent Information, 49,* A1–A3.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: System demonstrations* (pp. 55–60).

Mikolov, T., Chen, K., Corrado G., & Dean, J.(2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv: 1301.3781.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*(11), 39–41.

Moehrle, M. G., Walter, L., Geritz, A., & Müller, S. (2005). Patent- based inventor profiles as a basis for human resource decisions in research and development. *R&D Management, 35*(5), 513–524.

Park, H., Yoon, J., & Kim, K. (2012). Identifying patent infringement using SAO based semantic technological similarities. *Scientometrics, 90*(2), 515–529.

Park, H., Yoon, J., & Kim, K. (2013). Using function-based patent analysis to identify potential application areas of technology for technology transfer. *Expert Systems with Applications, 40*(13), 5260–5265.

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).

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*, pp. 11–18.

Phan, M. C., & Sun, A. (2018). CoNEREL: Collective information extraction in news articles. In *The 41st international ACM SIGIR conference on research & development in information retrieval* (pp. 1273–1276).

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*.

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

Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. arXiv preprint arXiv:cs/0306050.

Singh, S. (2018). Natural language processing for information extraction. arXiv preprint arXiv: 1807.02383.

Souili, A., Cavallucci, D., & Rousselot, F. (2015). Natural Language Processing (NLP): A solution for knowledge extraction from patent unstructured data. *Procedia Engineering, 131,* 635–643.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. I. (2012). BRAT: A web-based tool for NLP-assisted text annotation. In *Proceedings of the demonstrations at the 13th conference of the european chapter of the association for computational linguistics* (pp. 102–107).

Strzalkowski, T. (Ed.). (1999). *Natural language information retrieval*. Dordrecht: Kluwer.

Tsourikov, V., Batchilo, L., & Sovpel, I. (2000). Document semantic analysis/selection with knowledge creativity capability utilizing subject-action-object (SAO) structures (No. 6167370). Alexandria, VA: U. S. Patent and Trademark Office.

Wang, X., Qiu, P., Zhu, D., Mitkova, L., Lei, M., & Porter, A. (2015). Identification of technology development trends based on subject- action- object analysis: The case of dye-sensitized solar cells. *Technological Forecasting and Social Change, 98,* 24–46.

Wang, X., Ren, H., Chen, Y., Liu, Y., Qiao, Y., & Huang, Y. (2019). Measuring patent similarity with SAO semantic analysis. *Scientometrics, 121*(1), 1–23.

Wu, H. (2019). *Report of 2019 language & intelligence technique evaluation*. Baidu Corporation. http://tcci.ccf.org.cn/summit/2019/dlinfo/1101-wh.pdf, Accessed 24 Dec 2019.

Xu, S., An, X., Zhu, L., Zhang, Y., & Zhang, H. (2015). A CRF-based system for recognizing chemical entity mentions (CEMs) in biomedical literature. *Journal of Cheminformatics, 7*(Suppl 1), S11.

Xu, J., Guo, L., Jiang, J., Ge, B., & Li, M. (2019). A deep learning methodology for automatic extraction and discovery of technical intelligence. *Technological Forecasting and Social Change, 146*(9), 339–351.

Xu, S., Zhu, L., Qiao, X., Xue, C. (2009). A novel approach for measuring Chinese terms semantic similarity based on pairwise sequence alignment. In *Proceedings of the 5th international conference on semantics, knowledge and grid*, pp. 92–98.

Yang, C. B. (2012). Role of patent analysis in corporate R&D. *Pharmaceutical Patent Analyst, 1*(1), 5–7.

Yang, C., Huang, C., & Su, J. (2018). An improved SAO network-based method for technology trend analysis: A case study of graphene. *Journal of Informetrics, 12*(1), 271–286.

Yang, S., & Soo, V. (2012). Extract conceptual graphs from plain texts in patent claims. *Engineering Applications of Artificial Intelligence, 25*(4), 874–887.

Yang, C., Zhu, D., Bergmann, X., Zhang, Y., & Lu, J. (2017). Requirement-oriented core technological components' identification based on SAO analysis. *Scientometrics, 112*(2), 1229–1248.

Yoon, J., & Kim, K. (2012). An analysis of property–function based patent networks for strategic R&D planning in fast-moving industries: The case of silicon-based thin film solar cells. *Expert Systems with Applications, 39*(9), 7709–7717.

Yoon, J., Ko, N., Kim, J., Lee, J. M., Coh, B. Y., & Song, I. (2015). A function-based knowledge base for technology intelligence. *Industrial Engineering & Management Systems, 14*(1), 73–87.

Zeng, D., Liu, K., Chen, Y., & Zhao, J. (2015). Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1753–1762).

Zhang, L. (2016). *An integrated framework for patent analysis and mining (Doctoral dissertation)*. Miami, FL: Florida International University.

Zheng, S., Wang, F., Bao, H., Hao, Y., Zhou, P., & Xu, B. (2017). Joint extraction of entities and relations based on a novel tagging scheme. arXiv preprint arXiv:1706.05075.

Zhou, Y., Dong, F., Liu, Y., Li, Z., Du, J., & Zhang, L. (2020). Forecasting emerging technologies using data augmentation and deep learning. *Scientometrics, 122*(1), 1–29.