Journal of Informetrics 16 (2022) 101312

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

A deep learning based method benefiting from characteristics of patents for semantic relation classification

Liang Chen[a], Shuo Xu[b,*], Lijun Zhu[a], Jing Zhang[a], Guancan Yang[c], Haiyun Xu[d]

[a] Institute of Scientific and Technical Information of China, Beijing 100038, P.R. China
[b] College of Economics and Management, Beijing University of Technology, Beijing 100124, P.R. China
[c] School of Information Resource Management, Renmin University of China, Beijing 100872, P.R. China
[d] Business School, Shandong University of Technology, Zibo 255000, P.R. China

ARTICLE INFO

Keywords:
Semantic relation classification
Deep learning
Linguistic characteristics
Similarity measure
Patent analysis

ABSTRACT

The deep learning has become an important technique for semantic relation classification in patent texts. Previous studies just borrowed the relevant models from generic texts to patent texts while keeping structure of the models unchanged. Due to significant distinctions between patent texts and generic ones, this enables the performance of these models in the patent texts to be reduced dramatically. To highlight these distinct characteristics in patent texts, seven annotated corpora from different fields are comprehensively compared in terms of several indicators for linguistic characteristics. Then, a deep learning based method is proposed to benefit from these characteristics. Our method exploits the information from other similar entity pairs as well as that from the sentences mentioning a focal entity pair. The latter stems from the conventional practices, and the former from our meaningful observation: the stronger the connection between two entity pairs is, the more likely they belong to the same relation type. To measure quantitatively the connection between two entity pairs, a similarity indicator on the basis of association rules is raised. Extensive experiments on the corpora of TFH-2020 and ChemProt demonstrate that our method for semantic relation classification is capable of benefiting from characteristic of patent texts.

1. Introduction

As a valuable source of knowledge, patents are extremely important for technology opportunity discovery (Lee et al. 2019), invention protection (Park et al. 2012), technology trend analysis (Han et al. 2017) and so on. Though three types of information are contained in a patent document: (1) structured information, such as applicants and IPC (International Patent Classification) codes, (2) unstructured information, such as abstracts and claims, and (3) drawings describing the details of the invention, most novel technical intelligence is hidden in the unstructured information (Hu et al. 2018). Due to professionalism and complication of patent texts, technical intelligence is mainly obtained by expert reading (Yang 2012; Zhang 2016). Obviously, it 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.

It is well known that one of the vital techniques for computers to understand natural language is information extraction (Singh 2018). Last decade has witnessed significant progress in the domain of deep learning, and many successful cases from various

* Corresponding author.
E-mail addresses: 25565853@qq.com (L. Chen), xushuo@bjut.edu.cn (S. Xu), zhulj@istic.ac.cn (L. Zhu), janezhang@istic.ac.cn (J. Zhang), yanggc@ruc.edu.cn (G. Yang), xuhaiyunnemo@gmail.com (H. Xu).

https://doi.org/10.1016/j.joi.2022.101312
Received 16 January 2021; Received in revised form 25 June 2022; Accepted 29 June 2022
Available online 13 July 2022
1751-1577/

fields, such as computer vision (He et al. 2019) and speech recognition (Wang, Wang, & Lv 2019), are reported in the literature. The potential of deep learning in information extraction from patents has also been widely recognized in recent years (Chen et al., 2020a). With the extracted information, the inherent ambiguity in raw free texts can be reduced greatly so that many down-stream applications in IP domain can benefit from them. For example, Zhou et al. (2019) traced the system transformations and innovation pathways of an emerging technology with the information extracted from patent texts. An et al. (2018) monitored the technology trends of electric vehicle domain after 5 relationships were defined with the 13 identified prepositions.

However, compared to other cases, the potential of deep learning in patent information extraction hasn't been fully explored (Lupu 2017). In our opinion, main reasons are two-fold: (1) There are significant differences between patent documents and other types of documents, and even one can also observe significant differences between patent texts from different technical domains; (2) Many information-extraction tools embed a supervised machine learning model. This relies on the availability of a large amount of annotated instances from the target category of text resource. However, data annotation is quite laborious and time consuming, making enough annotated instances costly to obtain in practice for the rapid application of machine learning techniques in patent analysis.

Recently, three annotated corpora have been released in the literature, such as chemical patent corpus (Akhondi et al. 2014), patent corpus for chemical compound and drug named entity recognition (CHEMDNER-patents for short) (Pérez-Pérez et al. 2017) and thin film head patent corpus (TFH-2020 for short) (Chen et al. 2020a). It is expected that several customized information extraction tools specifically for patent documents will be developed on the basis of the cutting edge NLP (Natural Language Processing) techniques in the near future. For instance, Chen et al. (2020a) proposed a deep learning based framework for extracting semantic information from patents, and a competitive performance was achieved in terms of automation and accuracy. Though, the characteristics of patent texts remain largely unexploited in the literature, not to mention in-depth leverage of them for extracting valuable information from patents. Therefore, we identify the following research questions:

- To quantitatively measure the difference between patent texts from different technical domains as well as between patent texts and generic texts, so that the unique characteristics of patents can be highlighted.
- To figure out whether these unique characteristics of patents can be utilized to promote the state-of-art of semantic relation classification, a sub-task for information extraction, by customizing deep-learning models and what extent the customized models can benefit from these characteristics.
- To further investigate whether our customized deep-learning model can be generalized to other patent corpora. In this way, one can understand the adaptability and scope of applications of customized models.

The organization of the rest of this paper is as follows. After related work is briefly reviewed in Section 2, a deep-learning based method is put forward with the assistance of a similarity indicator to benefit from the linguistic characteristics of patent texts in Section 3. Then, extensive experiments are conducted on the corpora of TFH-2020 (Chen et al. 2020a) and ChemProt (Pérez-Pérez et al. 2017) to illustrate the advantages of this methodology in Section 4. The last section concludes this contribution with future study directions.

2. Related work

Before delving into more specifies, discussion of the literature pertinent to linguistic characteristics of patent texts, semantic relation classification for patents and deep learning methods for semantic relation classification is in order.

2.1. Linguistic characteristics of patent texts

According to requirements of patent laws, applicants must disclose the details of an invention to the public in exchange for a monopoly for a limited time (Idris 2004; Lupu et al. 2017). In the meanwhile, they have to reduce the level of disclosure to limit potential developments of the invention by their competitors (Fantoni et al. 2013). Such a paradox dramatically increases language complexity of patent texts, which enables patent documents to have linguistic characteristics different from those of generic texts.

The most significant characteristic of patent texts is compliance with the specification of the legal documents. That is to say, language statements should be in legal jargons. Thus, some synonyms in generic texts cannot be used interchangeably in patent documents, such as *consist of* and *comprise*. In patent context, the former implies an exhaustive enumeration and excludes any element, step, or ingredient not specified in the claims, whereas the latter commences an enumeration that is not necessarily exhaustive[1] and does not exclude additional elements[2]. In addition, in order to underline the novelty of the inventions and make them adaptable to changing circumstances, patent applicants may deliberately use different terminologies to express the same meaning, such as *light sensitive, photosensitive* and *photoreceptive* (Arinas 2012; Risch & Krestel 2019). Though the expressions in patent texts are usually more rigorous than those in generic texts (Hu et al. 2018; Xu et al. 2019), another popular phenomenon is polysemy, especially polysemic verbs describing actions, which severely jeopardize the subsequent analysis (Fantoni et al. 2013).

Furthermore, for the sake of protecting an invention, patent applications need to define the scope of the patent from others and cover as many variations as possible (Risch & Krestel 2019), but also increase the difficulty of their competitors to understand and

[1] https://www.epo.org/law-practice/legal-texts/html/caselaw/2019/e/clr_ii_a_6_2.htm
[2] https://www.bitlaw.com/source/mpep/2111_03.html

replicate the resulting invention (Fantoni et al. 2013). As a consequence, patent descriptions use many vague concepts. For example, a vacuum cleaner is mentioned as "Tornado generation method and apparatus" (Kanazashi & Yonedo 1998) and a document scanner is referred to as "light scanning apparatus" (Gomi et al. 2007). Another strategy pertinent to protection scope is the equivalents, which are alternative parts or steps that make the invention to work the same way and serve as the same purpose. For instance, if the invention is a handheld instrument that deposits a substance onto a surface for the purpose of writing, then the equivalents may include *ink, graphite, wax* and so on (Hunt et al. 2012). In order to make the patent defensible against rival patents and prevent the competitors from finding patentable alternatives of the patented invention, patent applicants often built a portfolio of patents by equivalent substitution, namely patent fence (Schneider 2008) to protect their technology.

It is these characteristics above that enable the sentences in patent documents to follow a more specific format than those in generic texts (Fantoni et al. 2013). It has been shown that property concepts are mainly expressed as "adjectives + nouns", such as *transparent window* and *anti-reflection film*, function concepts as "verbs + nouns", like *provide visibility* and *deposit nanoparticles* (Yoon & Kim, 2012a). Not only numerous multi-word concepts appear in patent documents, but also these concepts have plenty of variants with the same or similar meaning. Let's take the concept *thin film head* as an example. Its synonyms include *thin film magnetic head, thin film read/write head, thin film inductive head* and so on (Chen et al., 2020a). In addition, the long, complicated structure of sentence enables the entities and semantic relations to be more frequently mentioned in patent texts than in generic texts in most cases (cf. Table A2 in Appendix section). Xu et al (2019) observed that in contrast to scientific publications, the entities in patents are not mainly mentioned in terms of abbreviations or common names, but normative names of chemicals or chemical families with a defined structure. That is, the statements in patents are more formal than those in scholarly articles. As a matter of fact, rule-based information extraction models can benefit from these rigorous criteria. Chen et al (2020a) found that the performance of rule-based semantic relations extraction was 41.6% in term of precision, which was slightly lower than deep learning by 4.5%.

However, information extraction from patent documents still faces many challenges. As we all know, the core of a patent is composed of inventive principle to solve a problem and purpose that the invention achieves by solving it (Kim et al. 2018). As main expression of such domain-specific knowledge (Hu et al. 2018), technical jargons have the following three attributions. (1) Out-of-vocabulary**:** with the development of novel technologies, new unregistered technical jargons keep constantly emerging in patent texts. For example, *graph convolutional network*, and *multi-head attention* have appeared with the advances of artificial intelligence technique. (2) Sparsity**:** since patent documents concentrate on detailed description of inventions, many technical jargons only appear in a specific domain with a limited frequency while rarely in the other domains (Hu et al. 2018). (3) Combination: Arthur (2009) argued that a new technology is created by constructing, putting together, and assembling from previously existing technologies. When reflected in patent texts, new technical jargons are observed to source from combinations of old ones. For instance, a series of new *magnetic head* are usually expressed by assembling *magnetic head* with other words (Chen et al. 2020a), such as *perpendicular magnetic recording head, slant gap thin film magnetic head,* and *magneto-resistive magnetic head*.

2.2. Semantic relation classification for patent texts

As an important subtask of information extraction, semantic relation classification aims to assign a relationship label from a set of candidates to each relation mention (Jurafsky 2020). When extracting valuable information from patents, one can readily obtain the relation mentions by a variety of methods, such as SAO (Subject-Action-Object) method (Choi et al. 2012, 2013;Wang et al. 2015; Guo et al. 2016), template method (Dewulf 2011; Yoon & Kim 2012a, 2012b), but it is not trivial to determine the resulting type of semantic relation. It has been shown that as an important building block for patent knowledge graph (Deng et al. 2019) and technical semantic network (Sarica et al. 2020), semantic relation classification can benefit various down-stream tasks in patent analysis, like technology planning (Ki & Kim 2017), alternative technology identification (Wang, Ren, Chen, Liu, Qiao, & Huang 2019), technology opportunity discovery (An et al. 2018), patent map construction (Trappey et al. 2018), etc.

Hence, semantic relation classification has received increasing attention from both the bibliometrics field and the NLP field. Ever since pioneering work by Brin (1998), a variety of methods have been proposed for semantic relation classification, such as pipeline approach (Choi et al. 2012, 2013; Wang et al. 2015; Guo et al. 2016; Okamoto et al. 2017; Wang, Xu, & Zhu 2018) and joint approach (Zheng et al. 2017; Wang, Zhang, Che, & Liu 2018). However, due to distinct linguistic characteristics of patent documents, and time-consuming annotation and heavy workload for building a large-scale annotated patent corpus, the pipeline approach still dominates the semantic relation classification task for patent texts. Hence, this subsection mainly discusses the literature pertinent to the pipeline approach.

Just as its name implies, the pipeline approach usually consists of a series of processing steps, in which the output of previous step is fed into next step. On the whole, patent texts are first parsed to PoS (Part-of-Speech) sequences and syntactic dependencies with a general-purpose NLP tool, and then interested relation mentions are filtered, normalized and classified with the help of a variety of lexical resources and manually-curated rules, as shown in Fig. 1. In what follows, key modules in Fig. 1 will be reviewed one by one.

(1) Relation mention acquirement

To acquire relation mentions for further analysis, several well-known text-mining and NLP tools have been utilized in the literature. Wang et al. (2015) and Guo et al. (2016) used *Stanford CoreNLP* (Chen et al. 2014) to parse patent texts. With the assistance of *GoldFire* (former name Knowledgist2.5^{TM}) (Invention Machine Corporation, 2001), Choi et al. (2012, 2013) derived relation mention candidates. However, due to linguistic characteristics of patent texts, the general-purpose tools inevitably suffer from parsing inaccuracy (Burga et al. 2013; Fantoni et al. 2013; Wang et al. 2015). To our knowledge, neither open-source or commercial advanced text-mining model is catered specifically to take linguistic characteristics of patent documents into consideration until now.

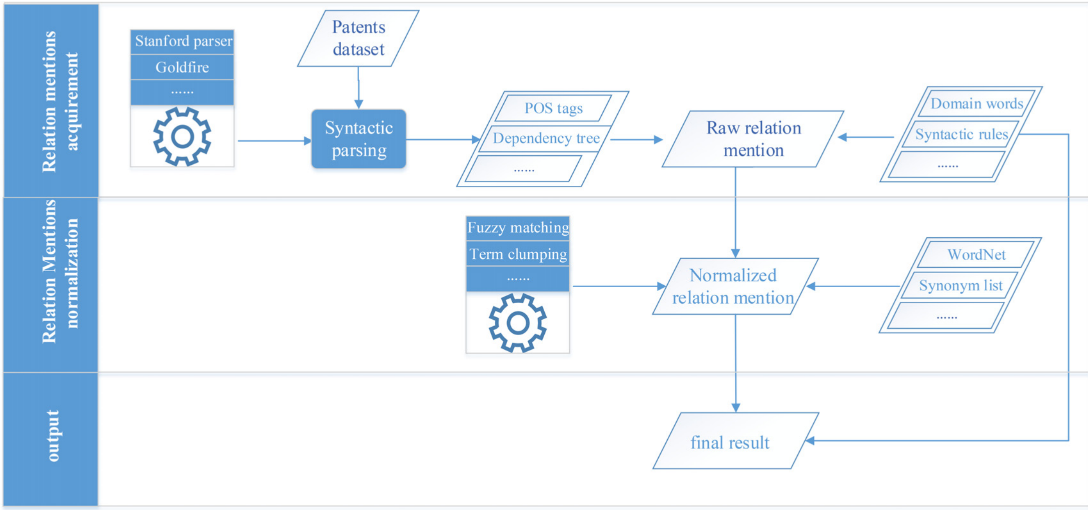


Fig. 1. General procedure of semantic relation classification in patent texts.

In order to filter interested relation mentions, mapping-rules should be built in advance and matched with parsed patent texts. A simple type of rule is word combination which is summarized from interested relation mentions (Yang et al. 2008). However, due to the aforementioned problems of vagueness and out-of-vocabulary (cf. Subsection 2.1), most studies devote to matching relation mentions with PoS tags combination and syntactic dependencies parsed from patent texts. Dewulf (2011) summarized that a function was mainly expressed in verbs and property concepts in adjectives. Yoon & Kim (2012a) further refined the PoS compositions of functions and attributes as "adjectives + nouns" and "verbs + nouns", and then five Stanford-type dependencies that can be grammatically represented by these PoS compositions were utilized to filter the relation mentions. As for the non-functional relations, An et al (2018) defined five semantic relations (*inclusion, objective, effect, process*, and *likeliness*), and used prepositions between these entities to determine their relationships.

(1) Relation mention normalization

Relation mention normalization aims to reduce word vagueness with the help of a public available or self-built vocabulary, knowledge base and similarity algorithms, thus to improve relation classification in patent texts. Bergmann et al. (2008) developed a domain-specific vocabulary based filter to standardize and minimize the highly differentiated idiom in the biotechnology field. The WordNet (Miller 1995), a large knowledge base of English, was utilized in Yoon et al. (2015) to collapse concepts semantically identical into common concepts by concept generalization. As for strongly domain-specific abbreviations not included in WordNet, Choi et al. (2012) took such abbreviations as sets of synonyms, namely *synsets* in the WordNet, and integrated them into the WordNet.

To alleviate the burden of the construction of lexical resources, several machine learning methods are adopted in the literature. Choi et al. (2012) employed a k-means clustering algorithm for vocabulary construction. In more details, the WordNet-based measurement is used to calculate the similarities between SAO structures, and then AOs and word phrases are further clustered into several groups for synonym folding. Instead of using lexical resource, VantagePoint provides a fuzzy matching function to find terms with similar structure in patents, thus to collapse terms semantically identical into representative one (Yang et al. 2018).

One can see that how to calculate semantic similarity between two interested terms plays an important role in the construction of lexical resources. To this end, Dao & Simpson (2005) proposed two classic algorithms which are widely used in follow-up research (Yoon & Kim 2011; Park & Yoon 2014; Huang et al. 2015; Xu et al. 2009; An et al. 2021). (1) The taxonomy in WordNet is treated as an undirected graph and then similarity of different terms is measured by the path length between them; (2) The normalized number of LCS (least common sub-summer) is taken as an indicator to measure the similarity between two terms, where the LCS refers to the summer that does not have any children that are also the sub-summer of two entities.

In addition, Tseng et al. (2007) used the statistical information of terms in a corpus for similarity calculation. Xu et al. (2009) argued that the difference of word order in the entities would affect the quality of corresponding similarity. As a remedy, they proposed an improved approach based on pairwise sequence alignment. Further, An et al. (2021) improved semantic analysis for assessing patent similarity on the basis of entities and semantic relations (functional and non-functional relations), in which the semantic direction of each sequence structure and the word order information of each component were considered simultaneously.

(1) Pre-defined types of relations

Due to the huge difference between patent texts in different fields, there is no consensus on the schema for relation types until now. In fact, when defining semantic relations, one should take application scenarios into consideration. Pérez-Pérez et al. (2017) proposed a schema containing 23 types of relations, such as *activator, agonist, cofactor*, for patents pertaining to biomedical science. As for thin

Table 1
Several schemas for semantic relations in patent texts.

Publication	Technological field	# of relation types	Relation type
Yang et al. (2008)	chemical mechanism polishing	8	contain, spatial, reference, functionality, utility, etc
Yang et al. (2012)	chemical mechanism polishing	4	verb, contain, prep, hasAttr
Choi et al. (2013)	Polymer electrolyte membrane techniques in proton exchange membrane fuel cells	3	Purpose function, effect type, partative type
Yoon et al. (2015)	all fields	25	object function, structural relation and attribute function containing 23 subclasses
Wang et al. (2015)	dye-sensitized solar cells	6	Inorganic dye, organic dye, photoanode, electrolyte, cell, electrode
Pérez-Pérez et al. (2017)	chemical biology	23	activator, agonist, antagonist, cofactor, agonist-activator, etc.
An et al. (2018)	electric vehicle	5	Inclusion, objective, effect, process, likeness
Chen et al. (2020)	thin film head techniques in hard disk drive	17	alias, part-of, measurement, operation, generating, instance-of, etc.

film head techniques in hard disk drive, Chen et al. (2020a) concentrated on the relations of *part-of, made-of, spatial relation* and so on in magnetic head. Several representative studies are illustrated in Table 1.

2.3. Deep learning methods for semantic relation classification

Instead of manually-curated features in conventional methods, deep learning based methods can provide an end-to-end paradigm to learn automatically discriminative features from the provided data for semantic relation classification. With the successful cases based on deep learning increase in many sub-tasks of NLP and bibliometrics, such as information retrieval (Guo et al. 2020), topic extraction (Zhang et al. 2018) and citation prediction (Abrishami & Aliakbary 2019), researchers began to explore the feasibility of deep learning technique in semantic relation classification task.

A variety of methods based on RNN (Recursive Neural Network) model were proposed to learn latent features on lexical and sentence level and yielded competitive performance on this task (Socher et al. 2012; Hashimoto et al. 2013). Due to the inherent drawback of RNN model in parallel computation, Zeng et al. (2015) replaced the RNN model with the CNN (Convolutional Neural Network) model, and vectorized the relative distance of the current word from a focal entity. It has been shown that this operationalization is effective, and has been adapted as an indispensable module by many follow-up approaches (Nguyen & Grishman 2015; Lin et al. 2016; Wang et al. 2016). In the meanwhile, the attention mechanism was widely applied to classify semantic relations (Lin et al. 2016; Zhou et al. 2016; Han et al. 2019). Wei et al. (2020) further took a more powerful neural network, namely Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2018), as the encoder to extract features and achieved state-of-the-art performance. Similarly, Lee & Hsiang (2019) built a pre-trained model, namely PatentBert by fine-tuning a released BERT-Base model on a patent classification task. As we all know, the vocabulary of the released BERT was built on Wikipedia and BookCorpus (Devlin et al., 2018). This is quite different from that on patent corpus and may cause the reduced performance of downstream tasks. As a remedy, Google trained a BERT model from scratch on over 100 million patent publications from the U.S. and other countries[3].

As increasing number of annotated patent corpora are released, the applications of deep learning on relation classification in patent texts have emerged. Chen et al. (2020a) employed a deep neural network to categorize semantic relations between a pair of entities in patent texts. In the following study (Chen et al. 2020b), they deeply discussed how to choose a suitable word embedding and deep learning model for this task. As a matter of fact, in previous studies, many deep learning models are directly incorporated to categorize semantic relations from patent texts, regardless of their distinct linguistic characteristics (cf. Subsection 2.1). Hence, compared to the performance on generic texts, the performance on patent documents is unsatisfactory. To our knowledge, it is still unknown that whether such characteristics can be utilized to promote the performance of deep learning models in semantic relation classification. Thereupon, this study wants to develop a novel deep learning based method so that this method can benefit from the distinct linguistic characteristics of patent texts.

3. Research framework and methodology

3.1. Unique characteristics of patent texts

To quantitatively illustrate the difference from the perspective of linguistic characteristics, this study collects seven annotated corpora covering three different categories: *news, encyclopedia* and *patent*. The news corpora consist of CoNLL-2003 (Sang et al. 2003) and NYT-2010 (Riedel et al. 2007), encyclopedia corpora include Wikigold (Balasuriya et al. 2009) and LIC-2019 (Wu 2019), and

[3] https://cloud.google.com/blog/products/ai-machine-learning/how-ai-improves-patent-analysis

patent corpora comprise CPC-2014 (Akhondi et al. 2014), ChemProt (Pérez-Pérez et al. 2017) and TFH-2020 (Chen et al. 2020a). To illustrate the difference between these corpora, several syntactic and lexical complexity indicators (Lu et al. 2019; Xu et al. 2021) are utilized here, as shown in Table A1 and A2 in Appendix section. Note that not all indicators are applicable for each corpus, since semantic relations are not labeled at all in CPC-2014, CoNLL-2003 and Wikigold, and LIC-2019 is not in English but in Chinese.

From Table A2 in Appendix section, three interesting phenomena can be observed: (1) There is no clear distinction between patent texts and generic texts in term of sentence length, which is different from the observation in Xu et al. (2021); (2) More entities and relations are mentioned in a single sentence of patent texts than that of generic texts in most cases; (3) The probability of co-words among entities in patent texts is higher than that in generic texts in most cases; (4) The TFH-2020 corpus shows clear distinction from the other patent datasets and all generic datasets in terms of 5 out of 7 indicators, such as number of entities per sentence, number of relations per sentence, entity length, proportion of n-gram entities, and entity association rate. To say it in another way, significant distinctions exist not only between patent texts and generic texts, but also between patent texts from different technical domains. Relatively speaking, the difference between patent texts and generic texts is easier to understand. As for the difference between patent texts from different technical domains, we argue that main reason is that plenty of sequences and chemical structures are mentioned in patent documents in the *chemical and biotechnology* field (Hunt et al. 2012), while the most frequent entities in the field of *hard disk drive* are of components, locations and functions (Chen et al. 2020a), as to describe inventions with different materials and mechanisms.

3.2. Research framework

In this subsection, we take the TFH-2020 corpus (Chen et al. 2020a) as our experimental dataset, and devote to exploiting its characteristics to improve the performance of sentence-level semantic relation classification. The TFH-2020 corpus contains 1,010 patent abstracts pertaining to *thin film head* technology in hard-disk collected from the USPTO (United States Patent and Trademark Office) database. To describe the technical content in this domain, Chen et al. (2020a) defined 17 types of entities (cf. Table A3 in Appendix section) and 15 types of semantic relations (cf. Table A5 in Appendix section) and annotated manually the resulting mentions. For clarity, the resulting examples are also provided for each type of entity and relation in Table A4 and A6 in Appendix section. We refer the readers to Chen et al. (2020a) for more details on this corpus.

On closer examination of the TFH-2020 corpus, one can find several interesting characteristics (cf. Table A2 in Appendix section) in terms of number of entities per sentence, number of mentions per entity, proportion of n-gram entities, and entity association rate. It is high entity association rate that our method for semantic relation classification benefits from. The entity association rate actually measures the possibility of establishing connection between entities in a corpus through co-word relation (cf. Table A1). For ease of understanding, two entities, namely *magnetic film* and *inductive thin film*, are taken as examples and they are connected since they share the common word *film*. As the entity association rate of TFH-2020 corpus is much higher than those of other corpora (cf. Table A2), the corresponding semantic relation triplets, viz. (entity1, relation type, entity2), are also much more closely related to each other. Given that the entities in a patent corpus are so closely related to one another, the relation type of a focal entity pair can be determined with the help of the types of other similar entity pairs. Therefore, a natural idea is to leverage the rich connections between relation triplets as well as the sentences mentioning these triplets to promote the performance of semantic relation classification.

According to our observation, the stronger the connection between two relation triplets is, the more likely they belong to the same relationship. Intuitively, this can be utilized to infer entity pair's relation type and further improve the performance of sentence-level relation classification. For this purpose, a new semantic relation classification model, as shown in Fig. 2, is put forward in this study. It is not difficult to see that our model consists of two sub-models: Bidirectional Gating Recurrent Unit-Hierarchical Attention Network (BiGRU-HAN) and Graph Convolutional Network (GCN). The former takes advantage of the features from the sentences mentioning an interested entity pair for relation classification, and the latter determines relation type by referring to the types of other similar relation triplets. In this way, our model takes simultaneously two groups of valuable information into consideration.

To answer our second question in Section 1, our combined model in Fig. 2 is extended to that in Fig. 3 in our implementation by utilizing the multi-input and multi-output characteristics of TensorFlow (Chollet 2017). Three sub-models are involved in Fig. 3: Weight GCN (WGCN for short), Weightless GCN (WLGCN for short), and BiGRU-HAN. In this way, apart from the output of BiGRU-HAN-WGCN, four more outputs are added to our method in Fig. 3: WLGCN output, WGCN output, BiGRU-HAN, and BiGRU-HAN-WLGCN output.

In fact, the advantage of the implementation in Fig. 3 is that it is very convenient to conduct a comparison analysis between our combined model and its sub-models by training each model independently and then put the validation results together to find their difference. As we all know, due to the randomness in network initialization and mini-batch creation, even for the same model its performance will vary every time after retrained. This is referred to as random difference in this paper. Thus, the performance difference between different models trained independently consists of model difference and random difference. To reduce the influence of random difference on the performance of the resulting model as much as possible, five outputs are generated on the same batch data from our implementation with shared parameterized models. The resulting code of our method can be accessed public freely[4].

It is noteworthy that since the relation type as the prediction target is an unknown variable in our combined model, the triplet (entity1, relation type, entity2) is converted into entity pair (entity1, entity2) and the triplet connection network is converted into entity pair connection network. Furthermore, before measuring the strength of entity pair connection, the connection between entities

[4] https://github.com/awesome-patent-mining/BiGRU-HAN-GCN

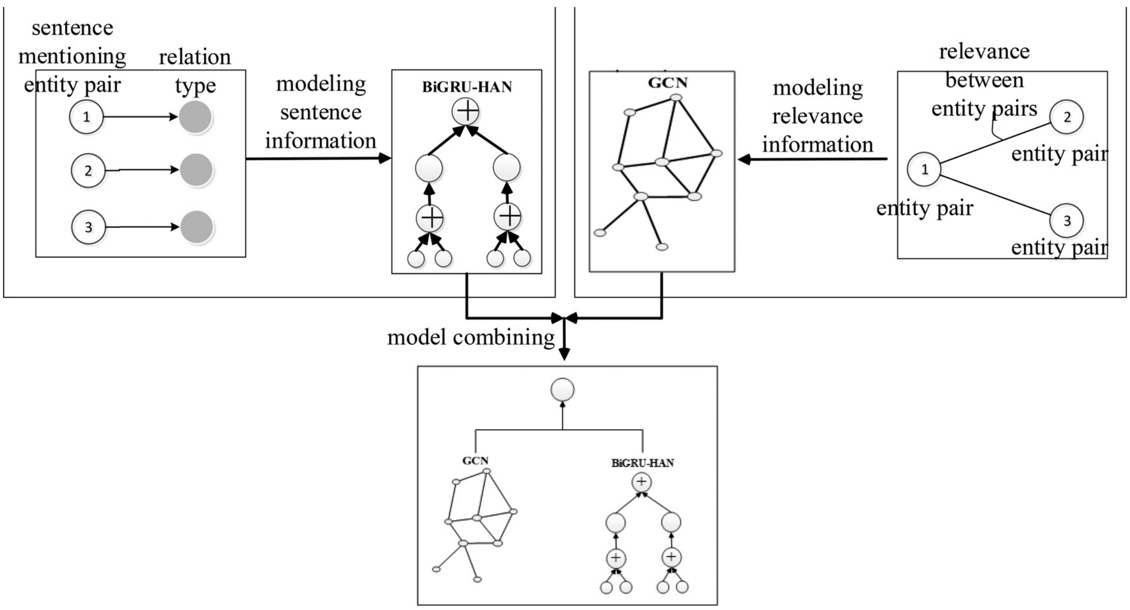


Fig. 2. The structure of our combined model for semantic relation classification

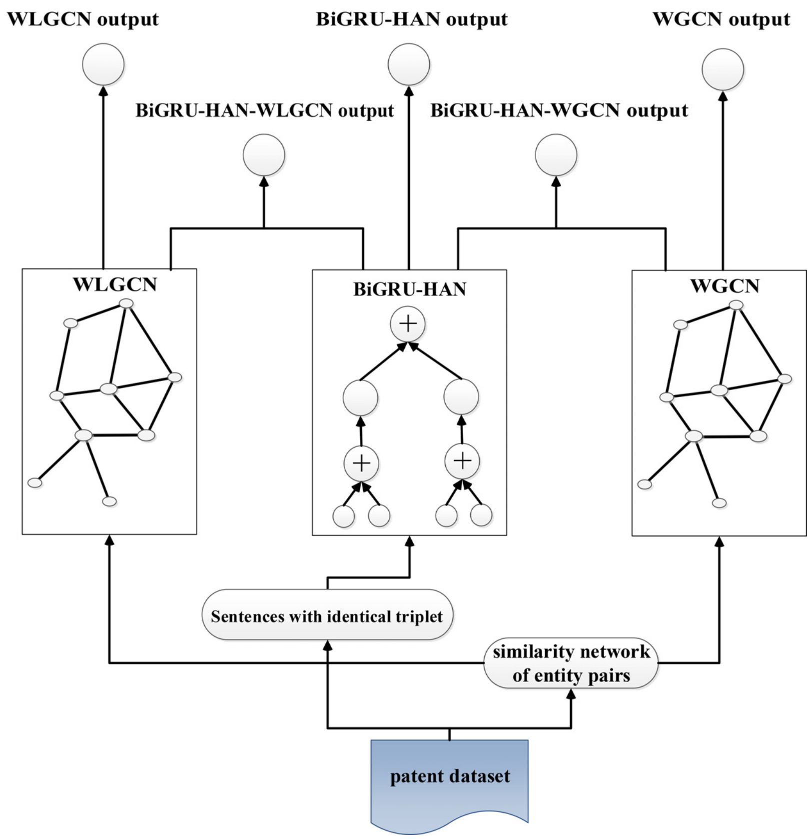
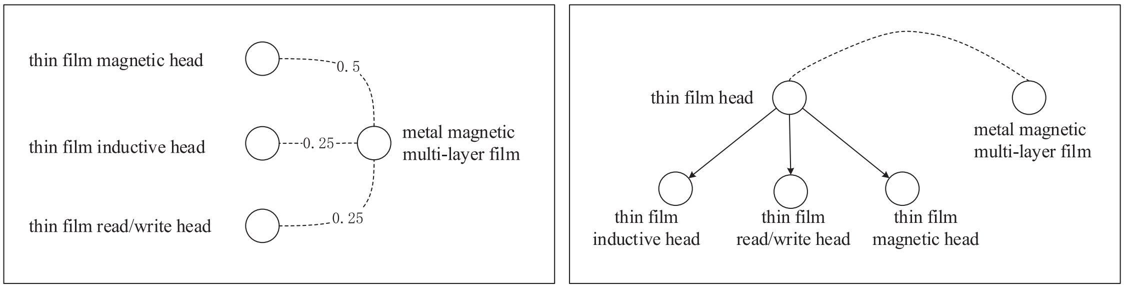


Fig. 3. The implementation of our combined model.



(a) cosine similarity (b) association rules method

Fig. 4. Different methods to measure the connection between entities

ID	Entity			
1	thin	film	magnetic	head
2	thin	film	inductive	head
3	thin	film	read/write	head

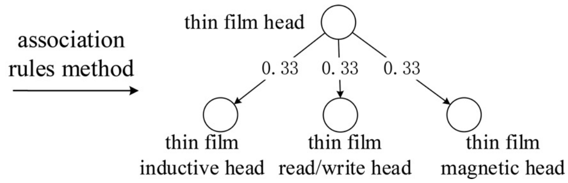


Fig. 5. A generic representation of entities found by association rules method

should be measured first. Though many conventional similar measures, such as cosine and Euclidean distance, can be used directly here, the performance is usually unsatisfactory in this case. In our opinion, the main reason is that these measures do not consider a special linguistic phenomenon in patent texts: most new entities are created by constructing, putting together, and assembling from previous entities. To validate our speculation, the entity *thin film magnetic head* is taken as an example. Many synonym variants appear in the TFH-2020 corpus, such as *thin film inductive head* and *thin film read/write head*. The cosine similarity between *metal magnetic multi-layer film* and *thin film magnetic head* is 0.5, while the similarities between *metal magnetic multi-layer film* and other variants are 0.25, as illustrated in Fig. 4(a). These unequal similarities are obviously counter-intuitive.

Given unavailable synonym vocabulary, to overcome this question, a novel similarity measure is developed in next subsection on the basis of association rules. Before delving into more specifies, the rational of our measure is clarified in the first place. The common part among these entity variants, *thin film head* in the example above, is discovered by association rules method as shown in Fig. 5. Then, only *thin film head* is used to measure the connection with *metal magnetic multi-layer film*, as shown in Fig. 4(b). Thus, the entity *metal magnetic multi-layer film* has a same similarity with *thin film magnetic head* and its variants. In addition, since association rules originate from the co-word relationship between entities over the corpus, our similarity measure can take contextual information into account when measuring the connection between entities. This is much more comprehensive than conventional measures that only involve matching score between text strings.

3.3. Similarity between entity pairs

There are two steps in calculating the connection (hereinafter referred to as similarity) between entity pairs, which will be described as follows.

(1) **Calculating similarities between entities**

Let *a* and *b* be two entities, and a' and b' be their maximal closed frequent sub-itemsets (Han et al. 2011) respectively:

Case 1: if a' is a subset of b' or vice versa, then

$$sim(a,b) = \begin{cases} 1 + \frac{\sigma(b')}{\sigma(a')},\ a' \subseteq b' \\ 1 + \frac{\sigma(a')}{\sigma(b')},\ b' \subseteq a' \end{cases} \quad (1)$$

where σ(·) indicates the support of the itemset.

Case 2: if a' is not a subset of b' nor vice versa, and c', the intersection of a' and b', is not empty, then

$$sim(a,b) = \frac{\sigma(a')}{\sigma(c')} + \frac{\sigma(b')}{\sigma(c')} \quad (2)$$

Case 3: if a' is not a subset of b' nor vice versa, and c', the intersection of a' and b', is empty, then

$$sim(a,b) = 0 \quad (3)$$

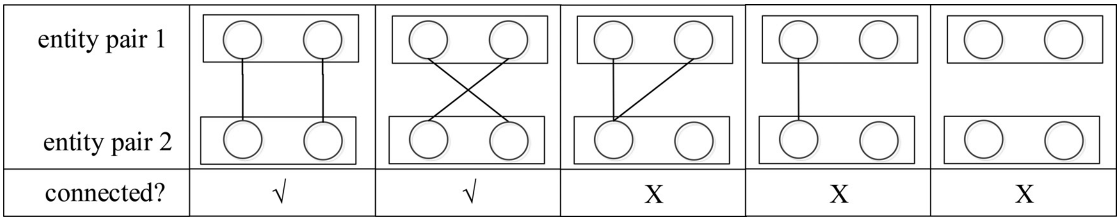


Fig. 6. Connections between entity pairs in different situations

(1) **Calculating similarities between entity pairs**

One can connect two entity pairs if each entity from one pair shares a word at least with each entity from the other pair. In this way, five situations may be encountered in real-world scenarios, as shown in Fig. 6. The link hints that there is at least one common word between the linked entities. Only the first two situations are allowable in this study. Of course, other decisions on the connection are possible, depending on the practical applications.

Let's take the following two entity pairs as an example:

Entity pair ep_1: (*magnetic film, thin film head*);

Entity pair ep_2: (*thin film magnetic head, inductive thin film*);

As can be seen, *magnetic film* from ep_1 and *thin film magnetic head* from ep_2 have two common words *magnetic* and *film*, and *thin film head* from ep_1 and *inductive thin film* from ep_2 share two words *thin* and *film*. Since each entity from ep_1 shares one word at least with each entity from ep_2, then one can connect ep_1 and ep_2.

When calculating the similarity between two connected entity pairs, since their relation types are unknown and entities from different entity pairs may are not aligned with each other, the resulting entity similarities are used to determine how to align the entities from two entity pairs, as illustrated in Eq. (4).

$$sim(A, B) = \mathrm{Max}\{\mathrm{sim}(a1,\ b1) + \mathrm{sim}(a2,\ b2),\ \mathrm{sim}(a1,\ b2) + \mathrm{sim}(a2,\ b1)\} \tag{4}$$

where $A = (a_1, a_2)$ and $B = (b_1, b_2)$ are the two entity pairs.

Example

To demonstrate the process of similarity calculation between entity pairs, the similarity between (*magnetic film, thin film magnetic head*) and (*inductive thin film, thin film head*) is taken as an example and calculated as follows.

At first, the entities in (*magnetic film, thin film magnetic head*) are paired with those in (*inductive thin film, thin film head*). The following tuples can be obtained:

T_1: *(magnetic film, inductive thin film),*
T_2: *(magnetic film, thin film head),*
T_3: *(thin film magnetic head, inductive thin film),*
T_4: *(thin film magnetic head, thin film head).*

The similarities between entities in these tuples will be calculated one by one. Let's take T_1 for example. The maximal closed frequent sub-itemsets of the two entities is {*film, magnetic*} with the support 19 and {*film, inductive, thin*} with the support 5, respectively. As {*film, magnetic*} is not a subset of {*film, inductive, thin*} and their intersection, {*film*} with the support 32, is not empty, Eq. (2) is applied to calculate the similarity as follows:

$$\begin{aligned} & sim(\ magnetic\ film,\ inductive\ thin\ film) \\ & = \frac{\sigma(\{film,\ magnetic\})}{\sigma(\{film\})} + \frac{\sigma(\ \{film,\ inductive,\ thin\})}{\sigma(\{film\})} \\ & = \frac{19}{32} + \frac{5}{32} \\ & = 0.75 \end{aligned}$$

Similarly, we can obtain the similarities between entities in T_2, T_3 and T_4 as 1.09, 0.65, and 1.75. Since $sim(T_1) + sim(T_4) > sim(T_2) + \mathrm{sim}(T_3)$, the former is chosen as the similarity between the entity pair (*magnetic film, thin film magnetic head*) and (*inductive thin film, thin film head*).

3.3. A combined model

Although the similarities between entity pairs have potential to improve the performance of sentence-level relation classification methods in theory, this is not actually in line with the assumption of sentence-level relation classification models. Specifically speaking, apart from the features in the sentence mentioning a relation instance, the relevant sentences also matter for classifying the relation instance. This point poses a big challenge for model construction and parameter estimation.

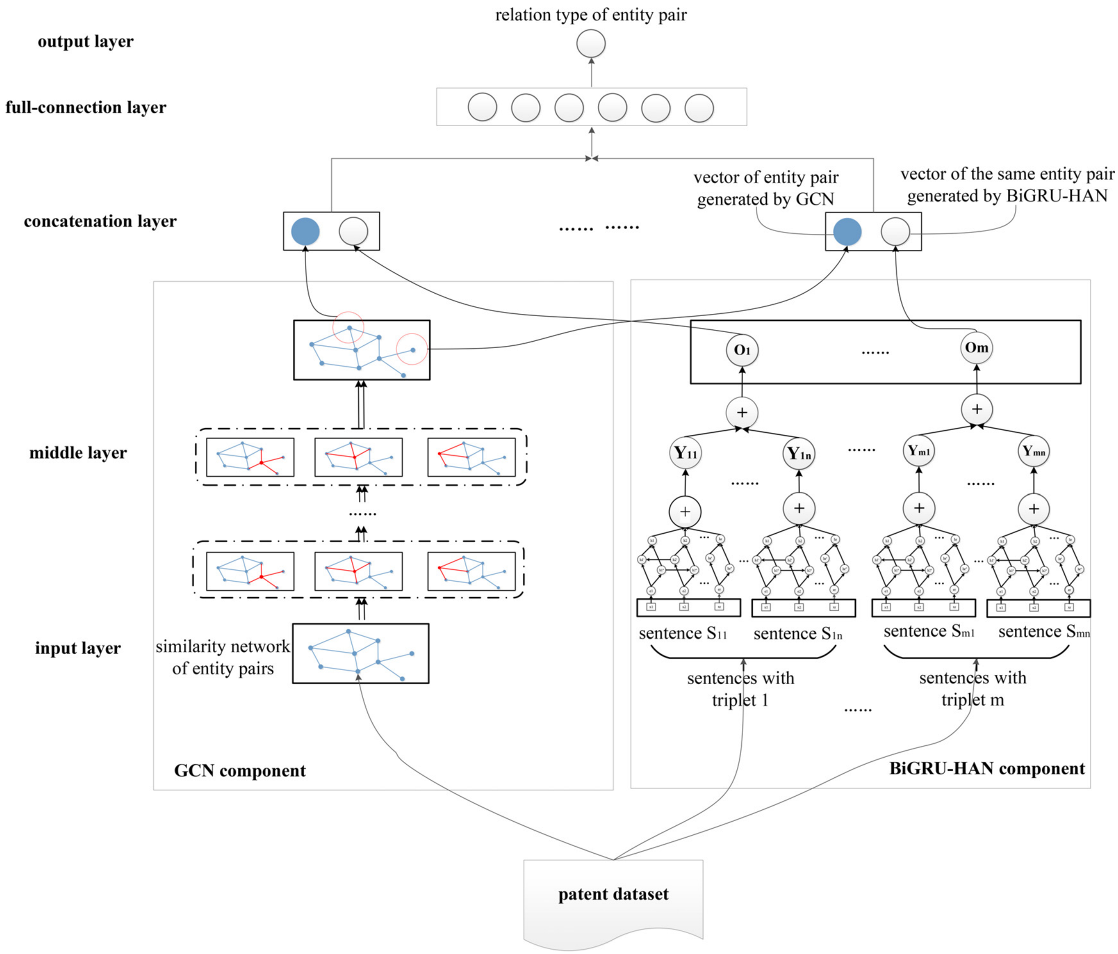


Fig. 7. The graph model representation of our combined model.

To meet the challenge, a combined model for relation classification, as shown in Fig. 7, is developed in this study. From Fig. 7, it is not difficult to see that our model is comprised of two sub-models: BiGRU-HAN (Han et al. 2019) and GCN (Kipf 2016). The BiGRU-HAN sub-model is made up of six layers (cf. right part in Fig. 7). The basic idea of BiGRU-HAN is to recognize the occurrence pattern of different semantic relations by a recurrent neural network, named as BiGRU, and then leverages a HAN (Hierarchical Attention Mechanism) consisting of a word-level attention layer and a sentence-level attention layer to further improve the model's prediction performance. The GCN is an extension of classic CNN (Convolutional Neural Network) (LeCun et al. 1989) to graph data aiming at representing a node as a low-dimensional vector preserving both network topology structure and node content information in it. As depicted in left part in Fig. 7, there are multiple graph convolutional layers in the GCN sub-model. In each layer, node representation is learned by propagating neighbor information in an iterative manner until a stable fixed point is reached. After several layers stacked together, the GCN sub-model has sufficient ability to extract high-level node representations and support downstream tasks.

In short, the BiGRU-HAN sub-model encodes a set of sentences mentioning an interested triplet and the entity pair within into a vector, and the GCN sub-model encodes the structure feature of the entity pair in a similarity network produced in Subsection 3.3 into a vector. Then, these two vectors corresponding to the same entity pair are concatenated in concatenation layer and input into full-connection layer to predict its relation type. However, the GCN is a full-batch model which requires calculating the representation of all the nodes in the graph for model training, while the BiGRU-HAN is a mini-batch model which divides the training dataset into mini-batches for training the model more efficiently. To synchronize data feeding of the two sub-models for training the combined model, the GCN is modified here to the mini-batch counterpart. In more details, training dataset are shuffled and split into mini-batches in the first place, and then for each mini-batch, only outputs corresponding to mini-batch are collected for back propagation algorithm running while other part of GCN are kept unchanged.

4. Experimental results and discussions

Two patent corpora, TFH-2020 (Chen et al. 2020a) and ChemProt (Pérez-Pérez et al., 2017), are utilized in this study to empirically evaluate our combined model (cf. Fig. 2). The first corpus consists of 1,010 patents pertaining to *thin film head* technology, and 15 entity types and 17 semantic relation types are involved. The second corpus comprises of 30,000 patents in the domain of *medicinal chemistry*, and 3 entity types and 23 semantic relation types are annotated manually. These two corpora come from different domains, and different scales of documents are involved in these two corpora. In this way, the robustness, scalability and adaptability of our combined model can be evaluated comprehensively.

4.1. Experimental setup

In this study, we choose TensorFlow (Abadi et al. 2016) as the deep learning framework and GloVe (Pennington et al. 2014) as the word embeddings. The maximal length of sentence and the number of epochs are fixed to 300 and 50 respectively. The learning rate for Adam optimizer is set to 0.001. In BiGRU-HAN sub-model, the sentence embedding size, word dimension and position dimension are 230, 50 and 5 respectively, the dropout probability is 0.5. In GCN sub-model, the numbers of layer and the node embedding size are 2 and 16 respectively. This setting is the most used configuration in deep learning based relation classification model training.

4.2. Evaluation measures

Herein two average metrics are employed to measure the overall performance of the five outputs: macro-average and weighted-average. In the macro-average metric, an interested measure index is calculated separately for each relationship, and then they are averaged. That is, the macro average metric does not take the imbalance of relation types into account. The weighted-average metric is very similar to macro-average metric, but a weight is attached to each relation type by the support (Macdonald et al. 2008). Formally, they are defined as follows.

$$\mathrm{B}_{macro}(\mathcal{M}) = \frac{1}{q}\sum_{j=1}^{q} B\left(TP_j, TN_j, FP_j, FN_j\right) \tag{6}$$

$$\mathrm{B}_{weighted}(\mathcal{M}) = \frac{support_j}{support_{all}}\sum_{j=1}^{q} B\left(TP_j, TN_j, FP_j, FN_j\right) \tag{7}$$

Here, $B(\cdot,\cdot,\cdot,\cdot)$ denotes one of the three most used classification measures: precision (P), recall (R), and F_1-value. $\mathcal{M}$ indicates the model for semantic relation classification, and q is the number of relation types. TP_j, TN_j, FP_j and FN_j represent respectively the number of true positive, true negative, false positive and false negative for the relation type j.

4.3. Experiment I: TFH-2020 corpus

4.3.1. Data preparation

According to Subsection 3.2, a similarity network of entity pairs is constructed from the TFH-2020 corpus. This network contains plenty of isolated vertices and small connected components. After components with the size less than 5 are removed, 9 connected components with 9,488 relation instances remain, in which the giant component has 9,437 relations. Then, the semantic relations are split randomly into a training set and a test set with a 7:3 ratio, which include 6,818 and 2,670 relations respectively.

As mentioned in Section 3, we assume that the greater similarity between two entity pairs indicates the higher probability that they belong to the same relationship. To demonstrate intuitively the usefulness of this hypothesis, all relation triplets of the TFH-2020 corpus are paired. Then, the similarity between two entity pairs from the paired relation triplets is computed. The distribution of pairs of entity pair (hereinafter referred to as PEP) with the percentile of similarities is shown in Fig. 7, in which column 1 enumerates all possible pairs of relation types in Table A3, such as (*operation, operation*), (*spatial relation, part-of*), etc. But, due to the limited space, for relation pairs with identical relation type, each relation type is denoted with the three-character abbreviation (cf. column 3 in Table A5). For example, (*operation, operation*) is written as *ope-ope*. As for the relation pairs with two different relation types, such as (*spatial relation, part-of*), the "not equal" is collectively used to refer to these relation pairs. Each cell in Fig. 8 indicates the number of PEP with the resulting semantic relationship pairs.

From Fig. 8, one can observe a positive correlation between the number of PEP with identical relation and the similarity except *cau-cau, spa-spa, mea-mea, oth-oth, ali-ali, ins-ins* and *gen-gen*. It is worth noting that very few instances are mentioned for the relation types: *mea, oth, ali, ins* and *gen* (cf. Table A5 in Appendix section). In the meanwhile, as for PEP with different relation types in *not equal* row, such positive correlation cannot be observed. To say it in another way, for most relation types in the TFH-2020 corpus, intuitively, our hypothesis holds.

4.3.2. Performance comparison

From the experimental results shown in Table 2, it can be seen that the combined models demonstrate great improvement over single sub-models by at least 2.6% in term of weighted-average F_1-value. This validates the usefulness of the similarities of entity pairs for semantic relation classification. On the other hand, one can observe that the performance of the WLGCN model is inferior to that of WGCN by 0.6% in term of the weighted-average F_1-value. This indicates that the network structure plays a more critical

relation types of entity pairs	percentile of similarities between entity pairs									
	100~90	90~80	80~70	70~60	60~50	50~40	40~30	30~20	20~10	10~0
ope-ope	72	34	18	10	17	7	14	1	9	1
com-com	63	3	3	15	7	2	6	11	9	5
att-attr	1563	450	229	261	200	211	165	110	118	131
for-for	76	48	50	48	63	98	105	69	45	57
mad-mad	100	135	509	397	311	151	66	57	32	12
man-man	148	141	74	84	94	51	40	39	21	13
par-par	7166	7900	7906	7001	5873	4780	3851	3975	3693	2956
pur-pur	183	274	215	134	157	81	117	86	120	134
mea-mea	2	0	0	0	0	0	0	0	0	0
oth-oth	0	0	0	0	0	0	0	0	0	0
ali-ali	1	0	1	0	0	0	0	0	0	0
ins-ins	2	8	1	0	6	3	0	6	7	0
gen-gen	7	4	10	2	2	6	6	2	1	10
cau-cau	224	261	290	295	162	185	209	260	266	331
spa-spa	4421	2684	2359	2277	3723	5225	6166	6387	7249	7104
not equal	32157	31915	31980	31631	32302	32056	31616	31403	30904	29960

Fig. 8. The distribution of entity pairs with the percentile of similarities.

Table 2
The performance of five models on the TFH-2020 corpus.

	macro-average (%)			weighted-average (%)		
	P	R	F_1	P	R	F_1
WGCN	19.1	18.0	17.5	39.4	46.0	41.0
WLGCN	22.4	18.1	17.9	39.4	45.4	39.6
BiGRU-HAN	42.0	40.5	41.0	63.0	63.4	63.2
BiGRU-HAN-WGCN	**45.8**	**43.5**	**44.3**	**66.3**	**66.7**	**66.4**
BiGRU-HAN-WLGCN	45.3	43.0	44.0	65.6	66.1	65.8

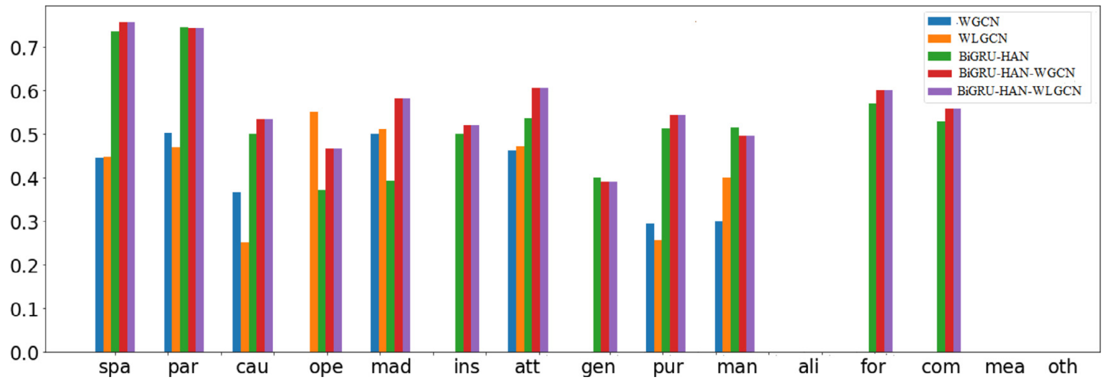


Fig. 9. The performance of five models on each relation type of the TFH-2020 corpus in term of precision.

role than similarity information for semantic relation classification. When such impact from the GCN sub-model is propagated to the combined model, similar phenomena can be observed for the BiGRU-HAN-WLGCN model and BiGRU-HAN-WLGCN model. This is an interesting finding for practical applications. Once the workload for calculating the edge weights is costly, the corresponding module can be removed from our method without much performance loss.

In order to highlight the performance of the five models on different types of relations, the precision, recall, and F_1-value for each type of relation are reported in Fig. 9-11. Though our combined model outperforms single sub-models for most types of relations, there are still several exceptions. From the perspective of precision, the WLGCN model performs better than our combined model on the relation type *operation*, and the BiGRU-HAN model is superior to our combined model on three relation types (*generating, part-of*,

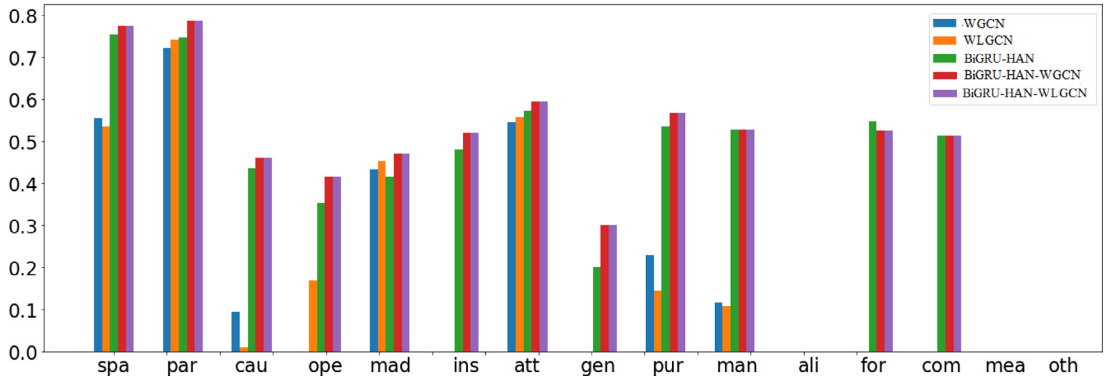


Fig. 10. The performance of five models on each relation type of the TFH-2020 corpus in term of recall.

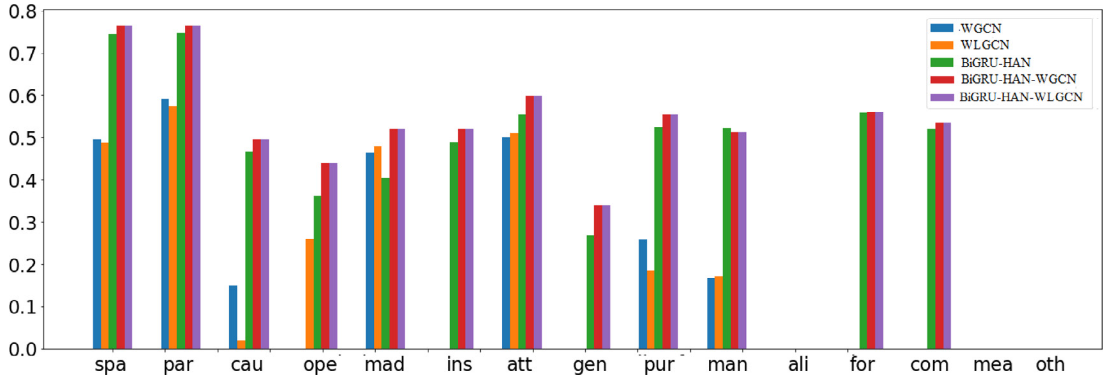


Fig. 11. The performance of five models on each relation type of the TFH-2020 corpus in term of F_1-value.

Table 3
Two-tailed statistical significance with 95% confidence interval by paired-samples t-test on TFH-2020 corpus

	macro-average F_1-value						weighted-average F_1-value				
	1	2	3	4	5		1	2	3	4	5
1	–	0.007	0.000	0.000	0.000	1	–	0.504	0.000	0.000	0.000
2	0.007	–	0.000	0.000	0.000	2	0.504	–	0.000	0.000	0.000
3	0.000	0.000	–	0.008	0.003	3	0.000	0.000	–	0.000	0.001
4	0.000	0.000	0.008	–	0.047	4	0.000	0.000	0.000	–	0.045
5	0.000	0.000	0.003	0.047	–	5	0.000	0.000	0.001	0.045	–

Note: 1-WGCN, 2-WLGCN, 3-BiGRU-HAN, 4-BiGRU-HAN-WGCN, 5-BiGRU-HAN-WLGCN

and *in-manner-of*). In addition, the performance of the BiGRU-HAN model is also better than that of our combined model on two relation types (*formation* and *in-manner-of*) in terms of recall and F_1-value, respectively.

In this way, it seems that the performance on five relation types (*operating, generating, part-of, formation*, and *in-manner-of*) does not conform to our expectation. As a matter fact, the support of *in-manner-of* is only 997 which accounts for 5.7% of the total relation mentions (cf. Table A5). In other words, for the minority relation type, the BiGRU-HAN model should be preferred to. However, as for the other four relation types, though the WLGCN model outperforms our combined model in term of precision or recall, it underperforms our combined model in term of recall or precision by a large margin. This results in a big superiority of our combined model in term of F_1-value.

From Fig. 8-11, one can observe an interesting phenomenon. The number of entity pairs does not correlate positively with the resulting similarities for the *attribution* and *spatial* relationships, but our combined model still performs better than the other counterparts. This indicates that the classification task for these relation types is still able to benefit from our combined model.

In addition, two-tailed paired-samples *t*-tests with 95% confidence interval are conducted amongst these five models (Xu, 2018), as reported in Table 3. One can see that the performance difference between our combined models (BiGRU-HAN-WGCN and BiGRU-HAN-WLGCN) and the other models is statistically significant in terms of macro-average and weighted-average F_1-values. Table 3 also illustrates that there is no statistically significant difference between the WGCN and WLGCN models.

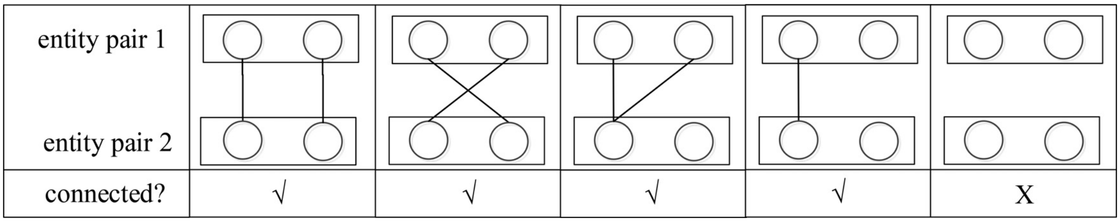


Fig. 12. Connections between entity pairs in different situations for the ChemProt corpus

Table 4
The performance of five models on the ChemProt corpus.

model name	macro-average (%)			weighted-average (%)		
	P	R	F_1	P	R	F_1
WGCN	27.9	26.9	26.5	43.9	48.7	45.2
WLGCN	29.8	24.8	24.7	44.1	47.3	43.3
BiGRU-HAN	63.6	62.3	62.3	69.5	69.0	69.0
BiGRU-HAN-WGCN	63.3	**64.4**	**63.6**	**71.5**	**70.9**	**71.0**
BiGRU-HAN-WLGCN	**63.7**	63.7	63.3	71.0	70.4	70.5

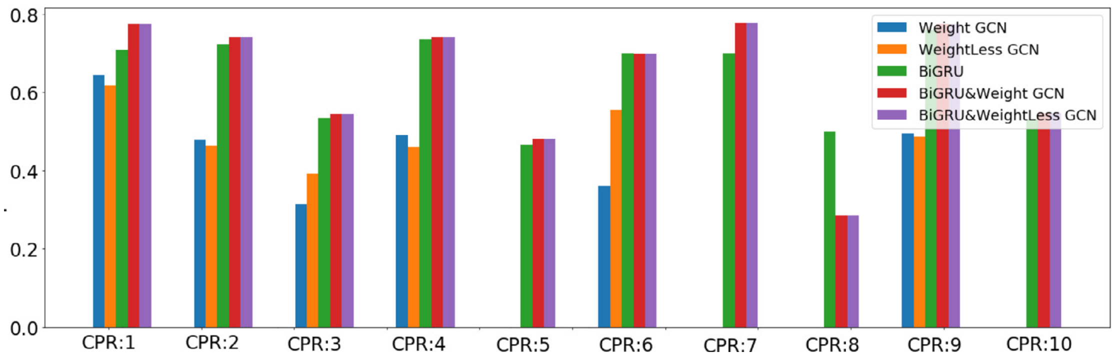


Fig. 13. The performance of five models on each relation type of the ChemProt corpus in term of precision

4.4. Experiment II: ChemProt corpus

4.4.1. Data preparation

Different from the TFH-2020 corpus, the ChemProt corpus (Pérez-Pérez et al., 2017) comes from *medicinal chemistry* domain. Therefore, chemical compounds, genes, proteins and their relations are annotated manually. More specifically, all entity mentions are grouped into three categories: CHEMICAL, GENE-N and GENE-Y. 23 relation types are involved in this corpus, such as ACTIVATOR, AGONIST, ANTAGONIST, and so on. To focus on the key relation types, all relation types are further grouped into 10 semantically related classes that share some underlying biological properties (Pérez-Pérez et al., 2017). These semantic classes are labeled as CPR:1, CPR:2, …, CPR:10, respectively.

From Table A2 in Appendix section, it is evident that most linguistic indicators of this corpus are much lower than that of the TFH-2020 corpus, especially in term of entity association rate. This indicates that the connections between entity pairs in this corpus are much less than those of the TFH-2020 corpus. That is to say, the similarity network is very sparse in term of the number of edges. Too sparse network is insufficient for training a deep learning model, so the establishment condition of connection between entity pairs is relaxed, as shown in Fig. 12. In the end, the resulting similarity network contains 17 connected components with 7,822 relation instances, in which the giant component has 7,781 relation instances. Similar to the TFH-2020 corpus, the semantic relations are split randomly into a training set and a test set with a 7:3 ratio, which include 5,585 and 2,237 relations respectively.

4.4.2. Performance comparison

The performance of the five models is shown in Table 4, and the precision, recall, and F_1-value for each type of relation are reported in Fig. 13-15. Again, our combined model outperforms the other models. But compared to the TFH-2020 corpus, the improvement gap is reduced in term of weighted-average F_1-value (2.6% vs. 1.5%). We argue that main reason is that the connection condition of two entity pairs is relaxed, which enables the connections of similarity network from this corpus not to be so closely related as those from the TFH-2020 corpus. Though, this noisy similarity network still promotes the performance of our combined model. This indicates that the similarity indicator raised in this paper is effective in improving semantic relation classification.

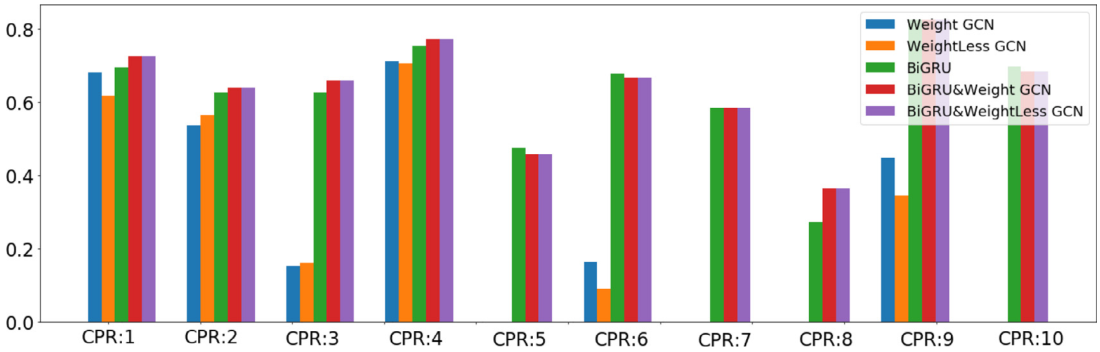


Fig. 14. The performance of five models on each relation type of the ChemProt corpus in term of recall.

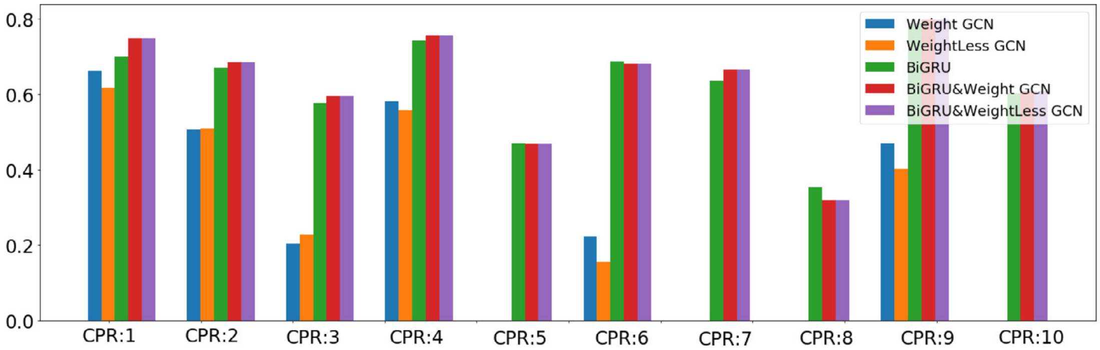


Fig. 15. The performance of five models on each relation type of the ChemProt corpus in term of F_1-value.

Table 5
Two-tailed statistical significance with 95% confidence interval by paired-samples *t*-test on the ChemProt corpus

macro-average F_1-value						weighted-average F_1-value					
	1	2	3	4	5		1	2	3	4	5
1	–	0.001	0.000	0.000	0.000	1	–	0.001	0.000	0.000	0.000
2	0.001	–	0.000	0.000	0.000	2	0.001	–	0.000	0.000	0.000
3	0.000	0.000	–	0.573	0.953	3	0.000	0.000	–	0.000	0.001
4	0.000	0.000	0.573	–	0.422	4	0.000	0.000	0.000	–	0.003
5	0.000	0.000	0.953	0.422	–	5	0.000	0.000	0.001	0.003	–

Note: 1-WGCN, 2-WLGCN, 3-BiGRU-HAN, 4-BiGRU-HAN-WGCN, 5-BiGRU-HAN-WLGCN

It is worth noting that our combined model achieves an average-weighted F_1-value of 71.0% on this coprus, which is significantly superior to that of 64.8% on the TFH-2020 corpus. In our opinion, main reasons are two-fold: (1) The number of entity types in this corpus is much less than that in the TFH-2020 corpus. Even though 23 relation types were annotated in the ChemProt corpus, they were grouped into 10 categories by the authors[5]. Due to the reduction of relation types, the difficulty in semantic relationship classification is alleviated to some extent. (2) The proportion of n-gram entities in the ChemProt corpus is much lower than that in the TFH-2020 corpus (cf. Table A2 in Appendix section). This indicates that most relation triplets from the former consist of uni-gram entities, which enables our model to determine the relation type with less uncertainty. Similar to the experiment I, we refer readers to the experimental results directory at https://github.com/awesome-patent-mining/BiGRU-HAN-GCN/ for more details.

Similarly, two-tailed paired-samples *t*-tests are also conducted with 95% confidence interval, depicted in Table 5. No significant differences are observed amongst BiGRU-HAN, BiGRU-HAN-WGCN and BiGRU-HAN-WLGCN models in term of macro-average F_1-value. But our combined models (BiGRU-HAN-WGCN and BiGRU-HAN-WLGCN) are obviously better than the counterparts. This again indicates that our proposed similarity indicator encodes important information for semantic relation classification.

5. Conclusions

As an important task to derive technical intelligence from patent texts, information extraction has drawn increasing attention in the bibliometrics field. Especially, with the rise of deep learning technique, a bunch of cutting edge NLP models are utilized to extract

[5] https://biocreative.bioinformatics.udel.edu/media/store/files/2017/ChemProt_Corpus.zip

valuable information from patent documents. However, most studies just transferred these models from generic texts to patent texts while keeping structure of the models unchanged. In fact, there are clear distinctions not only between patent text and generic text, but also between patent texts from different technical domains. This enables the performance of these models in patent texts to be reduced dramatically.

To deal with this problem, an important subtask of information extraction, namely semantic relation classification, is chosen as the focal task to exploit the possibility of leveraging the linguistic characteristics of patent texts for model improvement. In more details, a comprehensive comparison is conducted first among several annotated corpora from different fields to highlight unique characteristics of patent texts. From this comparison, a meaningful observation is obtained: the stronger the connection between two entity pairs is, the more likely they belong to the same relationship. To measure quantitatively the connection between two entity pairs, a similarity indicator is raised, and then a combined model by integrating the BiGRU-HAN and GCN sub-models is further developed for improving the performance of semantic relation classification.

In our combined model, two versions of the GCN sub-model, namely weighted GCN (WGCN) and weightless GCN (WLGCN), are constructed, and combined with the BiGRU-HAN sub-model through parameter sharing and multi-head output. The main goal of this combination strategy is to eliminate the random differences caused by independently training of different models and facilitate the following comparison analysis. Experimental results show that our combined models, BiGRU-HAN-WLGCN and BiGRU-HAN-WGCN, have achieved significant improvements of 3.2% and 2.6% in term of weighted-average F_1-value on the THF-2020 corpus, and 2.0% and 1.5% on the ChemProt corpus. This indicates that unique characteristics of patent texts can be used to promote the performance of semantic relation classification. More specifically, the structure of the connections between entity pairs as well as the degree of connections between them contributes greatly to the model improvement. It is worth mentioning that our method is not confined to patent texts. In fact, as long as a corpus has similar linguistic characteristics, our method should be applicable.

The following summarizes main contributions of this article:

- By quantitatively measuring the difference between patent texts from different technical domains as well as between patent texts and generic texts, the patent texts are found to have distinct characteristics from generic texts.
- To leverage the characteristics of patent documents, this study develops a deep learning based method for semantic relation classification, in which a similarity indicator based on association rules is raised to measure the connection between entity pairs.
- The superior performance of our method illustrates that the structure of the connections between entity pairs encodes important information for semantic relation classification.

Though, this study is subject to the following limitations. (1) According to our operationalization of the connection between entity pairs, the similarity network only contains about 50% entity pairs. Hence, how to construct a similarity network with more coverage is a necessary work for patent relation classification. (2) Several linguistic characteristics are observed in the patent documents, but the high entity association rate is just exploited in this work. Hence, how to exploit the potential of other characteristics also deserves further study.

Acknowledgements

This research received the financial support from National Natural Science Foundation of China under grant numbers of 71704169 and 72074014, 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 1

Tables A1–A6

Table A1
Indicators measuring linguistic characteristics.

Indicator	Description	Formula	Notation
Sentence length	Avg. number of words in a sentence	$L_{avg} = \frac{\sum_i^N L_i}{N}$	*N* indicates the number of sentences. L_i, SE_i and SR_i denote respectively the number of words, entities, and relations in the *i*-th sentence.
Number of entities per sentence	Avg. number of entities in a sentence	$SE_{avg} = \frac{\sum_i^N SE_i}{N}$	
Number of relations per sentence	Avg. number of relations in a sentence	$SR_{avg} = \frac{\sum_i^N SR_i}{N}$	
Entity length	Avg. number of words in an entity	$EW_{avg} = \frac{\sum_i^{NE} EW_i}{NE}$	*NE* and *RE* indicate respectively the number of entities and relations mentioned in a corpus, and *NE_distinct* and *RE_distinct* denote respectively the number of the entities and relations after de-duplication. EW_i is the number of words in the *i*-th entity, and NE_{ngram} indicates the number of entities consisting of multiple words in a corpus. $NE\_associated_i$ represents the number of unique entities that share one word at least with the *i*-th entity.
Number of mentions per entity	Avg. number of an entity mentioned in a corpus	$ER = \frac{NE}{NE\_distinct}$	
Number of mentions per relation	Avg. number of a relation mentioned in a corpus	$RR = \frac{RE}{RE\_distinct}$	
Proportion of n-gram entities	proportion of the entities consisting of multiple words	$EP_{ngram} = \frac{NE_{ngram}}{NE}$	
Entity association rate	Avg. association rate between different entities	$EA = \frac{100 * \sum_i^{NE\_distinct} NE\_associated_i}{NE\_distinct^2}$	

Table A2
Statistics on linguistic characteristics for several annotated datasets.

	Sentence length	Number of entities per sentence	Number of relations per sentence	Entity length	Number of mentions per entity	Number of mentions per relation	Proportion of n-gram entities (%)	Entity association
Rate								
CPC-2014 (EN)	23.3	2.5	—	1.4	5.3	—	25.7	1.6
ChemProt (EN)	21.9	2.4	0.6	1.3	3.7	4.73	19.3	0.3
TFH-2020 (EN)	30.7	6.1	4.3	2.3	2.8	1.2	75.5	7.6
CoNLL-2003 (EN)	14.6	1.7	—	1.5	33.3	—	37.6	0.06
Wikigold (EN)	23.0	2.1	—	1.8	5.1	—	50.4	0.6
NYTC (EN)	40.6	2.2	0.4	1.5	13.5	8.0	44.1	0.04
LIC-2019 (CN)	—	3.0	2.1	—	2.5	1.3	—	—

Table A3
The specification of entity types and the distribution of resulting instance in TFH-2020

	Entity Type	Abbreviation	Comment	# of entities	Proportion (%)
1	physical flow	Phy	substance that flows freely	140	0.6
2	information flow	Inf	information data	198	0.8
3	energy flow	Ene	entity relevant to energy	1,505	5.8
4	Measurement	Mea	method of measuring something	148	0.6
5	Value	Val	numerical amount	373	1.4
6	Location	Loc	place or position	2,450	9.5
7	State	Sta	particular condition at a specific time	40	0.2
8	Effect	Eff	change caused an innovation	694	2.7
9	Function	Fun	manufacturing technique or activity	1,455	5.6
10	Shape	Sha	the external form or outline of something	1,037	4.0
11	Component	Com	a part or element of a machine	12,452	48.3
12	Attribution	Att	a quality or feature of something	1,833	7.1
13	Consequence	Con	The result caused by something or activity	124	0.5
14	System	Sys	a set of things working together as a whole	958	3.7
15	Material	Mat	the matter from which a thing is made	1,643	6.3
16	scientific concept	Sci	terminology used in scientific theory	666	2.6
17	Other	Oth	Not belongs to the above entity types	66	0.3

Table A4
The examples for entity types in TFH-2020

Type	Example
physical flow	The etchant solution has a suitable solvent additive such as glycerol or methyl cellulose
information flow	A camera using a film having a magnetic surface for recording magnetic data thereon
energy flow	Conductor is utilized for producing writing flux in magnetic yoke
Measurement	The curing step takes place at the substrate temperature less than 200.degree
Value	The curing step takes place at the substrate temperature less than 200.degree
Location	The legs are thinner near the pole tip than in the back gap region
State	The MR elements are biased to operate in a magnetically unsaturated mode
Effect	Magnetic disk system permits accurate alignment of magnetic head with spaced tracks
Function	A magnetic head having highly efficient write and read functions is thereby obtained
Shape	Recess is filled with non-magnetic material such as glass
Component	A pole face of yoke is adjacent edge of element remote from surface
Attribution	A pole face of yoke is adjacent edge of element remote from surface
Consequence	This prevents the slider substrate from electrostatic damage
System	A digital recording system utilizing a magnetoresistive transducer in a magnetic recording head
Material	Interlayer may comprise material such as Ta
scientific concept	Peak intensity ratio represents an amount hydrophilic radical
Other	Pressure distribution across air bearing surface is substantially symmetrical side

Table A5
The specification of relation types and the distribution of resulting instance in TFH-2020

	Relation Type	Abbreviation	Comment	# of relations	Proportion (%)
1	spatial relation	Spa	specify how one entity is located in relation to others	3,444	19.8
2	part-of	Par	the ownership between two entities	4,431	25.4
3	causative relation	Cau	one entity operates as a cause of the other entity	1,490	8.6
4	operation	Ope	specify the relation between an activity and its object	852	4.9
5	made-of	Mad	one entity is the material for making the other entity	457	2.6
6	instance-of	Ins	the relation between a class and its instance	233	1.3
7	attribution	Att	one entity is an attribution of the other entity	2,575	14.8
8	generating	Gen	one entity generates another entity	222	1.3
9	purpose	Pur	relation between reason/result	1,497	8.6
10	in-manner-of	Man	do something in certain way	997	5.7
11	alias	Ali	one entity is also known under another entity's name	99	0.7
12	Formation	For	an entity acts as a role of the other entity	539	3.1
13	comparison	Com	compare one entity to the other	253	1.5
14	measurement	Mea	one entity acts as a way to measure the other entity	290	1.7
15	other	Oth	not belongs to the above types	33	0.2

Table A6
The examples for relation types in TFH-2020

Type	Example
spatial relation	Gap spacer material is then deposited on the film knife-edge
part-of	a magnetic head has a magnetoresistive element
causative relation	Pressure pad carried another arm of spring urges film into contact with head
Operation	Heat treatment improves the (100) orientation
made-of	The thin film head includes a substrate of electrically insulative material
instance-of	At least one of the magnetic layer is a free layer
Attribution	The thin film has very high heat resistance of remaining stable at 700.degree
Generating	Buffer layer resistor create impedance that noise introduced to head from disk of drive
Purpose	conductor is utilized for producing writing flux in magnetic yoke
in-manner-of	The linear array is angled at a skew angle
Alias	The bias structure includes an antiferromagnetic layer AFM
Formation	Windings are joined at end to form center tapped winding
Comparison	First end is closer to recording media use than second end
Measurement	This provides a relative permeance of at least 1000
other	Then, MR resistance estimate during polishing step is calculated from S value and K value

.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., . . . Zheng, X. (2016). Tensorflow: A system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation* (pp. 265–283).

Abrishami, A., & Aliakbary, S. (2019). Predicting citation counts based on deep neural network learning techniques. *Journal of Informetrics, 13*(2), 485–499.

... & Akhondi, S. A., Klenner, A. G., Tyrchan, C., Manchala, A. K., Boppana, K., Lowe, D., & Muresan, S. (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.

An, X., Li, J., Xu, S., Chen, L., & Sun, W. (2021). An improved patent similarity measurement based on entities and semantic relations. *Journal of Informetrics, 15*(2), Article 101135.

Arinas, I. (2012). How vague can your patent be? Vagueness strategies in US patents. *HERMES-Journal of Language and Communication in Business, 48*, 55–74.

Arthur, W. B. (2009). *The nature of technology: What it is and how it evolves*. New York, NY: Simon and Schuster.

Balasuriya, D., Ringland, N., Nothman, J., Murphy, T., & Curran, J. R. (2009). Named entity recognition inWikipedia. In *Proceedings of the 2009 Workshop on the People's Web Meets NLP* (pp. 10–18). ACL-IJCNLP 2009.

Bergmann, I., Butzke, D., Walter, L., Fuerste, J. P., & Erdmann, V. A. (2008). In *Evaluating the risk of patent infringement by means of semantic patent analysis: the case of DNA chips: 38* (pp. 550–562). R& D Management.

Brin, S. (1998). Extracting patterns and relations from the world wide web. In *International Workshop on The World Wide Web and Databases* (pp. 172–183).

Burga, A., Codina, J., Ferraro, G., Saggion, H., & Wanner, L. (2013). The challenge of syntactic dependency parsing adaptation for the patent domain. *ESSLLI-13 workshop on Extrinsic Parse Improvement*.

Chen, D., & Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 740–750).

Chen, L., Xu, S., Zhu, L., Zhang, J., Lei, X., & Yang, G. (2020a). A deep learning based method for extracting semantic information from patent documents. *Scientometrics, 125*(1), 289–312.

Chen, L., Xu, S., Shang, W., Wang, Z., Wei, C., & Xu, H. (2020b). What is special about patent information extraction? In *In the 1st Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents at the JCDL 2020* (pp. 63–72).

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

Chollet, F. (2017). *Deep learning with Python*. Shelter Island, NY: Manning Publications Co.

Dao, T. N., & Simpson, T. (2005). Measuring similarity between sentences. http://www.codeproject.com/KB/string/semanticsimilaritywordnet.aspx.

Deng, W., Huang, X., & Zhu, P. (2019). Facilitating Technology Transfer by Patent Knowledge Graph. In *Proceedings of the 52nd Hawaii International Conference on System Sciences* (pp. 1256–1264).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

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.

Fantoni, G., Apreda, R., Dell'Orletta, F., et al. (2013). Automatic extraction of function–behaviour–state information from patents. *Advanced Engineering Informatics, 27*(3), 317–334.

Gomi, A., Nomura, Y., et al. (2007). *Light scanning apparatus and method to prevent damage to an oscillation mirror, reducing its amplitude, in an abnormal control condition via a detection signal outputted to a controller even though the source still emits light* (No. 7557976). Alexandria, VA: U. S. Patent and Trademark Office.

... & Guo, J., Fan, Y., Pang, L., Yang, L., Ai, Q., Zamani, H., & Cheng, X. (2020). A deep look into neural ranking models for information retrieval. *Information Processing & Management, 57*(6), Article 102067.

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, 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.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Waltham, MA: Elsevier.

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.

Hashimoto, K., Miwa, M., Tsuruoka, Y., & Chikayama, T. (2013). Simple customization of recursive neural networks for semantic relation classification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1372–1376).

He, K., Girshick, R., & Dollár, P. (2019). Rethinking imagenet pre-training. In *Proceedings of the IEEE international conference on computer vision* (pp. 4918–4927).

Hu, J., Li, S., Yao, Y., Yu, L., Yang, G., & Hu, J. (2018). Patent keyword extraction algorithm based on distributed representation for patent classification. *Entropy, 20*(2), 104–122.

Huang, L., Shang, L., Wang, K., Porter, A. L., & Zhang, Y. (2015). Identifying target for technology mergers and acquisitions using patent information and semantic analysis. In *2015 Portland International Conference on Management of Engineering and Technology* (pp. 2313–2321).

Hunt, D., Nguyen, L., & Rodgers, M. (2012). *Patent searching: Tools & techniques*. Hoboken, NJ: John Wiley & Sons.

Idris, K. (2004). *WIPO Intellectual Property Handbook: Policy, Law, and use*. Geneva: WIPO publication.

Invention Machine Corporation. (2001). Knowledgist 2.5-Product Description http://www.triz.ch/KN25Prodesc.doc. Accessed 11 October 2021.

Jurafsky, D., Martin, J. (2020). Speech and language processing (the 3nd edition draft). https://web.stanford.edu/~jurafsky/slp3/. Accessed 11 October 2021.

Kanazashi, T, & Yonedo, K. (1998). *Tornado generation method and apparatus. (No. 6082387)*. Alexandria, VA: U. S. Patent and Trademark Office.

Ki, W., & Kim, K. (2017). Generating information relation matrix using semantic patent mining for technology planning: a case of nano-sensor. *IEEE Access, 5*, 26783–26797.

Kim, H., Joung, J., & Kim, K. (2018). Semi-automatic extraction of technological causality from patents. *Computers & Industrial Engineering, 115*, 532–542.

Kipf, T. N., & Welling, M. (2016). *arXiv preprint*.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation, 1*(4), 541–551.

Lee, J. S., & Hsiang, J. (2019). Patentbert: Patent classification with fine-tuning a pre-trained bert model. arXiv preprint arXiv:1906.02124.

Lin, Y., Shen, S., Liu, Z., Luan, H., & Sun, M. (2016). Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 2124–2133). (Volume 1: Long Papers).

Lu, C., Bu, Y., Wang, J., Ding, Y., Torvik, V., Schnaars, M., & Zhang, C. (2019). Examining scientific writing styles from the perspective of linguistic complexity. *Journal of the Association for Information Science and Technology, 70*(5), 462–475.

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

Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., & White, R. W. (2008). *Advances in information retrieval: 30th European conference on IR research*. Berlin: Springer.

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

Nguyen, T. H., & Grishman, R. (2015). Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing* (pp. 39–48).

Okamoto, M., Shan, Z., & Orihara, R. (2017). Applying information extraction for patent structure analysis. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 989–992).

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

Park, I., & Yoon, B. (2014). A semantic analysis approach for identifying patent infringement based on a product–patent map. *Technology Analysis & Strategic Management, 26*(8), 855–874.

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).

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

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.

Schneider, C. (2008). Fences and competition in patent races. *International Journal of Industrial Organization, 26*(6), 1348–1364.

Sarica, S., Luo, J., & Wood, K. L. (2020). TechNet: Technology semantic network based on patent data. *Expert Systems with Applications, 142*, Article 112995.

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

Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1201–1211).

Trappey, A. J., Trappey, C. V., Govindarajan, U. H., & Jhuang, A. C. (2018). Construction and validation of an ontology-based technology function matrix: technology mining of cyber physical system patent portfolios. *World Patent Information, 55*, 19–24.

Tseng, Y. H., Lin, C. J., & Lin, Y. I. (2007). Text mining techniques for patent analysis. *Information Processing & Management, 43*(5), 1216–1247.

Wang, D., Wang, X., & Lv, S. (2019). An overview of end-to-end automatic speech recognition. *Symmetry, 11*(8), 1018.

Wang, L., Cao, Z., De Melo, G., & Liu, Z. (2016). Relation classification via multi-level attention cnns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 1298–1307). (Volume 1: Long Papers).

Wang, S., Zhang, Y., Che, W., & Liu, T. (2018). Joint extraction of entities and relations based on a novel graph scheme. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (pp. 4461–4467).

Wang, X., Qiu, P., Zhu, D., Mitkova, L., Lei, M., & Porter, A. L. (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.

Wang, Z., Xu, S., & Zhu, L. (2018). Semantic relation extraction aware of n-gram features from unstructured biomedical text. *Journal of Biomedical Informatics, 86*, 59–70.

Wei, Z., Su, J., Wang, Y., Tian, Y., & Chang, Y. (2020). A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1476–1488).

Wu, H. (2019). *Report of 2019 language & Intelligence technique evaluation*. Baidu Corporation http://tcci.ccf.org.cn/summit/2019/dlinfo/1101-wh.pdf Accessed 11 October 2021.

Xu, S. (2018). Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science, 44*(1), 48–59.

Xu, S., Zhai, D., Wang, F., An, X., Pang, H., & Sun, Y. (2019). A novel method for topic linkages between scientific publications and patents. *Journal of the Association for Information Science and Technology, 70*(9), 1026–1042.

Xu, S., Li, L., An, X., Hao, L., & Yang, G (2021). An approach for detecting the commonality and specialty between scientific publications and patents. *Scientometrics, 126*(9), 7445–7475.

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

Yang, S. Y., Lin, S. Y., Lin, S. N., Lee, C. F., Cheng, S. L., & Soo, V. W. (2008). Automatic extraction of semantic relations from patent claims. *IJEBM, 6*(1), 45–54.

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

Yoon, J., & Kim, K. (2011). Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks. *Scientometrics, 88*(1), 213–228.

Yoon, J., & Kim, K. (2012a). 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., & Kim, K. (2012b). TrendPerceptor: A property-function based technology intelligence system for identifying technology trends from patents. *Expert system with application, 39*(3), 2927–2938.

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.

Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., & Zhang, G. (2018). Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics, 12*(4), 1099–1117.

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, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 207–212). (Volume 2: Short Papers).

Zhou, X., Huang, L., Porter, A., & Vicente-Gomila, J. M. (2019). Tracing the system transformations and innovation pathways of an emerging technology: Solid lipid nanoparticles. *Technological Forecasting and Social Change, 146*, 785–794.