IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 21, 2018, accepted July 30, 2018, date of publication August 3, 2018, date of current version October 12, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2862885

Gastroenterology Ontology Construction Using Synonym Identification and Relation Extraction

YING SHEN[1], YALIANG LI[2], YANG DENG[1], JIN ZHANG[1], MIN YANG[3], JINSONG CHEN[4], SHANGCHUN SI[1], AND KAI LEI[1]

[1]Shenzhen Key Lab for Information Centric Networking & Blockchain Technology, School of Electronics and Computer Engineering, Peking University, Shenzhen 518055, China
[2]Tencent Medical AI Lab, Palo Alto, CA 94301, USA
[3]SIAT, Chinese Academy of Sciences, Beijing 100871, China
[4]Shenzhen Maternal and Child Healthcare Hospital, Shenzhen 518040, China

Corresponding author: Kai Lei (leik@pkusz.edu.cn)

This work was financially supported by the National Natural Science Foundation of China (Grant 61602013), and the Shenzhen Key Fundamental Research Projects (Grant JCYJ20150301543307 11 and JCYJ20170818091546869).

ABSTRACT Ontology plays an increasingly important role in knowledge management and the semantic Web. However, ontology cannot perform well in realistic diagnosis reasoning unless it contains timely and accurate medical information and its individual items display all attributes of the categories they belong to. In this paper, we present a method that extracts synonyms along with concepts and their relationships for gastroenterology ontology construction. Specifically, we reuse the existing ontology as the basis for ontology completion. In addition, we conduct synonym identification through a combined application of global context features, local context features, and medical-specific features, and incorporate dependency information into deep neural networks for relation extraction. The extracted information is merged for ontology completion. Experimental results demonstrate that the proposed synonym identification and relation extraction method achieves the best performance compared with state-of-the-art methods and also builds a more complete ontology compared with existing gastroenterology disease ontologies. Our results are reproducible, and we will release the source code and ontology of this work after publication: https://github.com/shenyingpku/gastrointestinal_owl

INDEX TERMS Artificial neural networks, data acquisition, knowledge representation, machine learning, text mining.

I. INTRODUCTION

Constructing large-scale ontologies is important and useful for many real-world applications, such as medical decision support, precision medicine, and drug discovery. Ontologies provide a structured representation of the concepts of a domain of knowledge as well as the relations among them [1].

Various ontology construction approaches have recently been proposed. Ontology construction approaches can be classified in accordance with the type of knowledge resource, i.e., fully structured text [2], such as database, dictionary and ontology; semi-structured text [3], such as HTML and XML; or unstructured text, such as plain text in book, journal, and the webpage. Some general ontologies (e.g., Freebase [4] and ProBase [5]) are available, but most applications require a specific domain ontology to depict terms and relationships in that domain.

Despite the effectiveness of previous studies, medical ontology construction in the real world remains challenging. (i) First, medical synonym identification plays a crucial role in ontology construction. A single medical concept may have different ways of expression. In this context, synonym identification greatly facilitates entity alignment. Despite their usefulness, Chinese linguistic information and medical-specific knowledge, which play crucial roles in semantic comprehension, have been underexplored in recent work on Chinese medical synonym identification [4]. (ii) In addition, relation extraction (RE) is significant in robust knowledge extraction from unstructured texts in the medical domain and servers as an intermediate step in medical ontology construction. However, sentences in the medical domain are usually longer than those in the general domain [7]. Thus, the distances between target entities tend to be longer, which poses a difficulty for

capturing the relations between entities that are far apart. Moreover, labeled relation examples are often insufficient due to the high labeling cost.

To alleviate these limitations, we propose an approach that extracts synonyms along with concepts and their relationships for ontology construction. Specifically, we first reuse the existing ontology as the basis for ontology completion. Then, we conduct synonym identification through the combined application of global context features, local context features, and medical-specific features. Furthermore, we explore dependency information and incorporate this information into deep neural networks for relation extraction. Finally, the extracted information is merged to complete the proposed ontology. The proposed approach is implemented, and the performance of the proposed technique is evaluated.

The main contributions of our approach are fourfold:

1) We present a methodology for automatic ontology construction based on the existing ontology and for ontology knowledge completion via a synonym identification and relation extraction approach. We demonstrate the viability of producing a high-quality ontology.

2) We propose a synonym identification method through the combined application of linguistic features and medical-specific features, thus overcoming the limitations of dataset sparseness.

3) To tackle long-distance relation extraction in the medical field, we incorporate dependency information into deep neural networks to shorten the distance between medical entities and capture their relations.

4) There is no ontology for gastrointestinal diagnosis in the Chinese domain; therefore, the ontology we have constructed and released for this field may play a practical role in medical system applications in China.

In the rest of this paper, we review related work in Section 2. In Section 3, we describe our method of ontology construction in detail. Section 4 presents the comparison of the experimental results. Finally, we report our conclusion in Section 5.

II. RELATED WORK

Text-based ontology construction involves updating the existing ontology or mapping the existing ontologies to generate a large ontology. For instance, Alani [8] adopt ontology mapping and merging technology to perform automatic ontology construction from existing ontology. Li *et al.* [9] present an approach to merge ontology from various sources by using Formal Concept Analysis to extract formal semantics. Abinaya and Sumathi [10] utilizes WordNet-based semantic measures to identify similarities between concepts and terms from various ontologies and then merge ontologies. The difficulty of the first two methods lies in the need for human intervention to maintain the accuracy of the generated ontology. The problem of the last method is that it relies solely on the word-level information, yet it is clear that sentence representations that characterize the attributes of classes and instances are also important for merging ontologies.

To ease the automatic ontology construction process, this study proposes an approach that extracts synonyms along with concepts and their relationships. There are many approaches for performing synonym identification, e.g., lexical pattern, Online encyclopedia and search engine-based methods. Lexical pattern-based approaches require a good understanding of linguistics to propose rules and patterns [11]. Wikipedia-based methods have achieved the state-of-the-art performance because Wikipedia is rich in (semi-) structured knowledge, which is beneficial for synonym identification [12]. Many studies have attempted to use search engine-based methods on a large-scale search log data, which can be considered as a useful corpus for information mining [13]. Nevertheless, the Online encyclopedia and search engine-based methods ignore the information included in the plain text (e.g., morphological and lexical features). Support vector machines (SVMs) can be used for synonym identification [14]. Assisted by SVM, one can fully utilize the semantic information captured and present by the word embedding model, and capitalize on the unique radical and pronunciation characteristics of Chinese characters to identify synonyms.

Current text-based methods for automatic synonym identification be divided into three categories: knowledge-based, supervised, and unsupervised methods. Knowledge-based methods mainly use information from external knowledge sources or corpora of text [15]. The accuracy of semantic similarity measures relies heavily on the degree of knowledge completeness and structural sparseness of the adopted knowledge base. Supervised methods [16] adopt machine learning algorithms to assign concepts/terms to instances containing the ambiguous word. However, it is usually labor intensive and time-consuming to adopt the supervised methods since training data is needed be created for each target word to be disambiguated. In this study, we apply unsupervised methods [17] that adopt the distributional characteristics of a corpus to compute the semantic similarity. Several studies have attempted to explore the text corpus information with unsupervised methods. In general, there are two types of context information included in a text corpus, i.e., global context and local context. Global context and its topical information are explored by topic models, e.g., Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Analysis (PLSA), to learn the topic structure from the text. Local contexts can be trained by word embeddings, e.g., Word2Vec, to capture the semantic rules in the text. As a distributed representation of words, the basic idea of word embedding is to convert a word into a vector that is then projected into a low-dimensional vector space, such that similar vectors in the same space share higher relevance [18].

For relation extraction (RE), an inevitable drawback of supervised methods is that the data they use require human annotation and labeling, which is time-consuming and difficult to employ with a large corpus [19]. To alleviate this limitation, distant supervision methods have been widely explored in previous works [20]. Inspired by recent successes of deep learning in natural language processing, the

majority of studies have employed deep neural networks, e.g., convolutional neural networks (CNNs) or recurrent neural networks (RNNs), to realize distantly supervised relation extraction [21]. However, distant supervision methods cannot take full advantage of the linguistic information. Some recent studies leverage dependency information in deep neural network to make full use of data-driven models and semantic knowledge-driven models simultaneously. Socher *et al.* [22] use a RNN along the parse trees of sentences to conduct the sentiment analysis and relation classification. Liu *et al.* [23] develop a hybrid system that uses RNN to learn the subtrees and employs CNN to extract features on the shortest path. Cheng and Miyao [24] adopted bidirectional long short-term memory (Bi-LSTM) along dependency paths for temporal relation classification. In contrast to these methods, we experiment with attention-based BiGRU by considering (1) information on the whole dependency tree instead of only part of the tree (for instance, the shortest dependency path); (2) information between the entity and each word in the sentence as well as the relative relationship between the entity and each node in the dependency tree. Our model utilizes word- and sentence-level attention mechanisms and can achieve cross-sentence multi-relation extraction.

III. METHODOLOGY

A. EXISTING ONTOLOGY RE-UTILIZATION

Ontology construction may reuse the knowledge of existing ontologies and domain knowledge to give a commonly agreed understanding of a domain.

Automatic Chinese ontology construction is a difficult task due to the lack of a structured knowledge base or domain thesaurus. Our initial gastroenterology ontology was constructed based on Disease Ontology (DO).[1] The DO components relevant to gastroenterology, which include 14 types of disease, 341 classes, 346 inheritance relationships, and 61 equivalent relations, can be accessed through the subpath DOID:77 (disease → gastrointestinal system disease). All knowledge was translated from English to Chinese by a human translator.

The ontology knowledge completion was performed by completing the clinical text information in DO through synonym identification and relation extraction, which will be detailed in the following sections. Through entity matching, the extracted information is matched to the ontology components ''class name'' and ''alias'' that defined by the relationship ''hasExactSynonym'' in OWL. We consider the class not found in our generated ontology are new classes. For classes that already existed in the ontology, we compare and merge the information between the new input class and the existing class.

The newly added information was first stored in a MySQL database. Then, the owlready package[2] was adopted to convert the MySQL database to owl format. Protégé version 4.1 in OWL (Web Ontology Language) 2.0 was employed in this study. Some available plugins, such as OWLViz2 and OntoGraf, are used for the ontology generation.

[1] http://disease-ontology.org

[2] https://pythonhosted.org/Owlready/

B. ONTOLOGY KNOWLEDGE COMPLETION

1) SYNONYM IDENTIFICATION

Medical synonym identification is significant for building a high-quality medical ontology. Compared with high-frequency words, low-frequency words are trained with less information in a large-scale corpus, resulting in inaccurate entity knowledge presentation. Therefore, the application of context knowledge and linguistic information can reduce the requirement for knowledge completeness of the adopted dataset, thus improving the synonym identification of low-frequency words.

A text dataset contains two types of context information. The local context contributes to word sense disambiguation, while the global context is beneficial to provide useful topical information [25]. Following our previous work [26], this paper will improve the synonym identification by utilizing this two complementary context information collaboratively.

This study exploits the global context with the Normalized Google and Baidu Distance and discovers the local context with the cosine distance and edit distance both in English and Chinese, as well as the Chinese-specific radicals and pinyin edit distance. Furthermore, we explore the medicine-specific features that directly present information associated with the medical domain.

We employ SVM to identify whether a pairwise word are synonyms via the selected features.

a: LOCAL CONTEXT FEATURE EMBEDDING

Cosine Similarity: The similarity in the word vector space can be used to represent the semantic similarity of text. The similarity between two words can be measured by calculating the inner product between their vectors. The cosine similarity of the two-dimensional word vector A, B is given by:

$$\begin{aligned}\text{CosSim}\,(\text{A}, \text{B}) &= cos\,(A, B) = \frac{A \cdot B}{\|A\|\,\|B\|} \\ &= \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \end{aligned} \quad (1)$$

The cosine similarity is a judgment of orientation rather than scale: The cosine similarity of two vectors with the same orientation is 1; two vectors at 90° have a similarity of 0; and two completely opposite vectors have a similarity of -1, regardless of their scale. The larger the value, the more similar these two words are.

Radical Information: As hieroglyphs, Chinese characters are formed with radicals that have ideographic meanings. Chinese may have similar meanings when the radicals are the same [27]. For a pairwise Chinese word, frequently occurring radicals facilitate discrimination, such as 艹 (bacteria or drug), 疒 (disease or virus), 女 (female or gynecological diseases). This discrimination compensates for the lack

or inadequate representation of ideograms learned by other features.

To obtain the common radical information of two entities, the online Xinhua Chinese dictionary[3] was used to identify the radical of each Chinese character. Then, the number of common radicals, namely *commonRadicals* (A, B), is divided by the maximum length of A, B for normalization.

$$\mathrm{CR}\,(\mathrm{A},\,\mathrm{B}) = \frac{commonRadicals\,(A,B)}{maxLength\,(A,B)} \tag{2}$$

Edit Distance/Max Word Length: The edit distance between two strings is the minimum number of editing operations required (replace, insert and delete) to make the two strings identical. In general, the smaller the edit distance, the greater the similarity of the two strings. In programming functions, *editDist* (A, B) can be calculated via the edit distance algorithm. Therefore, the relative edit distance of the two words A and B is given by:

$$\mathrm{EditDist}\,(\mathrm{A},\,\mathrm{B}) = \frac{editDist\,(A,B)}{maxLength\,(A,B)} \tag{3}$$

Pinyin Edit Distance: Pinyin is unique to the Chinese language. Pinyin can eliminate differences in transliteration. For instance, the pinyin of 埃博拉病毒 and (both meaning Ebola virus) is the same. Same as radical information, we extract Pinyin information from the Xinhua Dictionary and ignore the changes in the four tones. The edit distance calculation method (Formula 3) is applied to evaluate the edit distance between two Pinyin sequences.

b: GLOBAL CONTEXT FEATURE EMBEDDING

Normalized Google and Baidu Distance: The Normalized Google Distance (NGD) [28] is adopted to measure the semantic similarity. Given a set of keywords, the NGD is about the number of hits returned by Google. A pairwise words that have similar meaning are often ''close'' in the NGD, while words with different meanings may be far apart. Given two search terms a and b, the NGD will be:

$$\mathrm{NGD}\,(a,b) = \frac{max\,\{\log f\,(a), \log f\,(b)\} - \log f\,(a,b)}{\log N - min\,\{\log f\,(a), \log f\,(b)\}} \tag{4}$$

where $f\,(a)$ and $f\,(b)$ are the hit numbers of terms a and b, respectively, $f\,(a, b)$ is the number of web pages where both a and b appear. N is the total number of Google search pages. $\log N$ is set to 10 in this study.

We can observe from formula 4 that NGD (a, b) is equal to 0 when terms a and b always occur simultaneously, indicating that a and b are very similar; NGD (a, b) is greater than or equal to 1 when terms a and b are very different; NGD (a, b) is infinite when terms a and b never appear together on the same Google page.

Baidu has the 2nd largest search engine in the world. Similar to the NGD, the same formula above is used in this study to calculate the Normalized Baidu Distance.

[3]http://xh.5156edu.com

c: MEDICAL FEATURE EMBEDDING

Taking medicine as an example, we make use of side effects, target proteins, mechanism of action and physiological effects to explore the medicine-specific features that can simplify their semantic representations for synonym identification.

Side Effect: The side effects of a drug d were obtained from the SIDER[4] database. Inverse Document Frequency (IDF) was applied to alleviate the impact of high-frequency terms and pay more attention to rarer ones:

$$\mathrm{IDF}\,(s, Drugs) = \log \frac{(|Drugs| + 1)}{(DF(s, Drugs) + 1} \tag{5}$$

where $Drugs$ is the set of all drugs, s is a side effect and $DF\,(s, Drugs)$ is the number of drugs with side effect s. The weighted side effect vector of a drug d is $sider(d)$, which consists of side effects extracted from SIDER. The value of element s of $sider(d)$ is $IDF(s, Drugs)$. The side effect-based relevance of two drugs d_1, d_2 is the cosine distance of the vectors $sider(d_1)$ and $sider(d_2)$.

Drug Target: The information about proteins targeted by a drug d was collected from DrugBank.[5] The target-based relevance of two drugs d_1, d_2 is defined as the cosine similarity of the IDF-weighted target protein vectors of two drugs, which are calculated like the IDF-weighted side effect vector.

Drug Mechanism and Physiological Effect: We collect both the mechanisms and physiological effect of a drug from NDF-RT.[6] The mechanism-based and physiological effect-based relevance of two drugs can be calculated by the cosine distance of the IDF-weighted mechanism vectors of the two drugs as mentioned in the previous paragraph.

2) ENTITY RELATION EXTRACTION

Relation extraction on a medical corpus is an important information extraction task in the medical domain and is the key step of ontology construction. However, in the medical domain, sentences are usually longer than those in the general domain. Thus, the distance between target entities tends to be longer, which poses a difficulty in capturing the relations between entities that are far apart.

To alleviate this challenge, we incorporate dependency information into deep neural networks. Dependency can provide more structural information on the tree parsed from sentences, thus enabling comparable performance with less training data and the capture of long-distance relations in the medical domain [29]. With distant supervision, data are automatically labeled by alignment with an existing knowledge graph.

a: INPUT REPRESENTATION

Given two medical entities (**en1** and **en2**) and a set of sentences (noted as *Sent*) containing both entities, we aim to

[4]sideeffects.embl.de/

[5]https://www.drugbank.ca/

[6]https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT

predict the relation $r \in R$ of the given entities, where R is the relation set.

Word Embeddings: The first layer of the network is the word embedding layer, which transforms words into representations that capture syntactic and semantic information about the words. Each word is converted into a real-valued vector. Therefore, the input to the next layer is a sequence of real-valued vectors. We choose Word2Vec to train word embeddings on a corpus containing Electronic medical records (EMRs) plain texts.

Dependency Feature Embeddings: To tackle the long-distance problem in the medical domain, dependency information is integrated into the deep neural network (i.e., attention-based BiGRU). By using dependency parsing, the linear sentence structure is transformed into a dense tree structure, and the long-distance relationship between two entities in a sentence can be better captured. In contrast to the commonly used shortest dependency path (SDP) models, we utilize information of the whole tree instead of only a part of the tree. Thus, the complete information obtained by the dependency tree can be used.

Dependency information is obtained from the hierarchical tree structure, including relative dependency features and dependency tags. The ***relative root features*** measure the relation between the current node and the root node. There are three types of relations here: the child node of the root, the root node itself, and others. The relative root features imply the relation between the current node and entity1 and entity2. There are four types of relations: the child node of entity1/entity2, the parent node of entity1/entity2, the entity node itself, and others. ***Dependency tags*** imply the tag of the current word to its parent node on the dependency tree.

We transform the dependency feature into real-valued vector representations for joint use with word embeddings. Then, the feature embeddings and word embeddings are concatenated.

b: ATTENTION-BASED BIDIRECTIONAL GRU

A gated recurrent unit (GRU) can allow each recurrent unit to adaptively capture dependencies of different time scales. GRUs are suitable for capturing relationships among sequential data. Bidirectional GRUs introduce a second layer to the unidirectional GRU networks. The hidden-to-hidden connections flow in opposite temporal order. The model is therefore able to exploit information from both the past and the future.

We use BiGRU as an encoder to read the source sentence via the embeddings of the words in the sentence. Referring to the selective attention mechanism [30], we apply the attention into the pooling layer to treat source representations as a memory and model the interaction between the decoder and the memory. The attention mechanism aims at recognizing which source sentences best represent the relation from the ones labeled through distant supervision. Finally, a softmax layer is used to calculate the probability predicted by the model of each class the input belongs to.

IV. EXPERIMENTAL RESULTS

A. SYNONYM IDENTIFICATION

1) DATASETS AND METRICS

To create training and test sets, we build a Chinese medical thesaurus semi-automatically. Terms are restricted to disease name and symptoms.

We mainly collect the synonym data from the medical-related pages of online encyclopedias via web crawlers. The Chinese online encyclopedia includes *Wikipedia*,[7] *Hudongpedia*[8] and *Baidu-baike*,[9] while the English corpus is derived from *DailyMed*[10] (up-to-date and accurate drug labels) and *WebMD*[11] (health information website).

We also obtain the synonym data from *a-plus encyclopedia*, *xunyiwenyao* (a popular crowdsourcing platform for doctor-patient communication), and *xiangya dictionary* (an authoritative online medical dictionary).

Accuracy ((TP+TN)/(TP+FN+FP+TN)), precision rate (TP/(TP+FP)), recall rate (TP/ (TP+FN)) and F1 score (2∗TP/(2∗TP+FP+FN)) were adopted as our evaluation metrics. TP, FP, TN and FN are True Positive, False Positive, True Negative and False Negative, respectively.

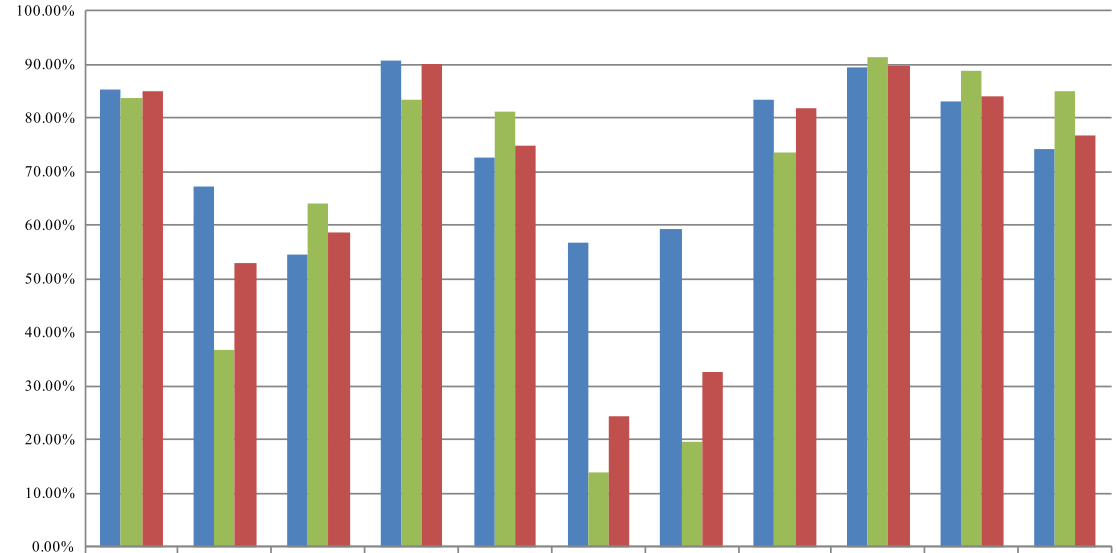
2) PERFORMANCE COMPARISON

The experimental results on our proposed dataset are summarized in Table 1. Several models are adopted for comparison: (1)-(3) PMI-related models aimed at estimating word similarity measures; (4)-(6) different word embedding models to learn sense embeddings for word similarity. To analyze the effectiveness of our model, we also report the ablation test in terms of discarding the local context features (w/o local), global context features (w/o global), or medical features (w/o medical).

We observe the following. (1) Our results have robust superiority over competitors and reveal the possibility of taking advantage of Chinese characters and medical-specific features to recognize synonyms in the medical field. (2) Generally, all features contribute in the similarity measure, enabling a larger performance boost to measure medical semantic similarity. It is within our expectation that integrating global and local features can effectively enrich knowledge representation. (3) Medical knowledge further enhances the knowledge representational learning of a specific domain. Even the basic model (w/o medical) achieves competitive results with these strong baselines, demonstrating the effectiveness of incorporating medical knowledge into the measurement of medical word similarity.

The experimental results for each feature are summarized in Figure 1. In general, Chinese character-related features (i.e., Chinese edit distance, Pinyin edit distance and radicals) perform best in Chinese medicine synonyms identification.

[7] http://download.wikipedia.com/zhwiki/latest/zhwiki-latest-pages-articles.xml.bz2
[8] http://fenlei.baike.com/%E5%8C%BB%E5%AD%A6/
[9] https://baike.baidu.com/science/medical
[10] https://dailymed.nlm.nih.gov/
[11] https://www.webmd.com/



	1.Cosine	2.SideEffect	3.DrugTarget	4.ChEdit Dist	5.EnEdit Dist	6.DrugMechanism	7.PhysiologicalEffect	10.Pinyin EditDist	11.Radical	12.Baidu Score	13.Google Score
precision	85.29%	67.22%	54.55%	90.85%	72.49%	56.77%	59.32%	83.52%	89.57%	83.10%	74.13%
Recall	83.74%	36.78%	64.00%	83.43%	81.19%	13.87%	19.58%	73.56%	91.38%	88.88%	84.94%
f1 score	85.11%	52.97%	58.57%	90.15%	74.77%	24.36%	32.58%	81.76%	89.79%	84.08%	76.72%

FIGURE 1. Results when using each feature exclusively.

TABLE 1. Result of synonym indetification.

Model	Accuracy
PMI (Terra and Clarke 2003) [30]	0.756
PMI-IR (Turney 2001) [31]	0.732
PPMI (Marek et al. 2011) [32]	0.755
GloVe (Iacobacci et al. 2015) [33]	0.637
Word2Vec (Handler 2014) [34]	0.722
Word2Vec (Hu et al. 2018) [35]	0.807
Our model	**0.897**
w/o local	0.819
w/o global	0.818
w/o medical	0.849

Cosine calculation is a reliable method for word similarity measure. It is worth noting that when searching Chinese medical terms, Baidu is superior to Google.

B. MEDICAL ENTITY RELATION EXTRACTION

1) DATA COLLECTION AND PREPARATION

The study was performed using more than 30,000 stroke EMRs. All of these EMRs between 2015 and 2016 are in the Chinese language.

For distant supervision relation extraction, the knowledge bases were aligned with sentences to permit automatic labeling. Several medical knowledge bases were used for the medical text labeling, including CN-Probase[12] (Chinese Concept Graph containing 17 million entities, 270 thousand concepts and 33 million relations), ICD-10[13] diagnosis codes (International Classification of Diseases), and CHPO[14] (The Chinese Human Phenotype Ontology Consortium). The relations included "disease_has_symptom", "disease_has_ risk-factor", "disease _has_complication", etc. Together with the relation "Others" (which implies there is no relation between two entities), there were 27 different types of relations. In total, 207,480 sentences were valid for the experiments. We randomly chose 20% of the data as testing data and treated the remaining data as training data. Held-out evaluation was employed in our model.

2) IMPLEMENTATION DETAILS

To train the neural network, cross-entropy was used to train the model. The Stanford dependency parser was chosen to extract the dependency features together with the POS features in the general field, whereas HanLP[15] was utilized in the medical domain.

The hidden size of BiGRU was 280. The dimension of word embeddings was set to 100, and all other feature embeddings were 10-dimensional. Other hyperparameters included a learning rate of 0.001, dropout probability of 0.5, and batch size of 100.

[12]http://kw.fudan.edu.cn/cnprobase
[13]http://www.icd10data.com/ICD10CM/Codes
[14]http://wiki.chinahpo.org/index.php
[15]https://github.com/hankcs/HanLP

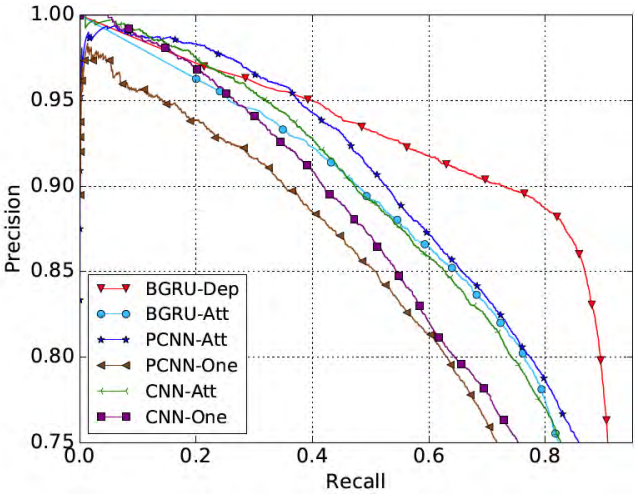


FIGURE 2. Aggregate precision/recall curves of CNN, CNN-One, CNN-Att, PCNN-One, PCNN-Att, BGRU-Att, BGRU-Dep in the medical domain.

3) PERFORMANCE COMPARISON

To demonstrate the effects of the dependency information, we empirically compared different distant supervision methods. The CNN model proposed in [37] and the PCNN model proposed in [38] were implemented as sentence encoders. For the two different types of CNN models, we compared the results of the sentence-level attention (-Att) proposed in [39] and the at-least-one multi-instance learning (-One) used in [38]. We denoted our dependency model as BGRU-Dep. Based on BGRU-Dep, and we proposed an ablation model (BGRU-Att) in terms of discarding dependency information.

Held-out evaluation was conducted, and the results are shown in Figure 2. Based on the results, we can conclude the following. (1) In general, our model (BGRU-Dep) achieves the best performance. Especially when the recall is larger than 0.4, our model outperforms the other models by a large margin. (2) Even the ablation model (BGRU-Att) with the attention mechanism achieves competitive results with these strong baselines. (3) BGRU-Dep consistently outperforms BGRU-Att, indicating the importance of dependency knowledge in relation extraction. (4) In the medical domain, our model does not suffer from the sharp decline in precision-recall curves observed in CNN and PCNN models at very low recall.

4) LESS TRAINING DATA

This section demonstrates that by adopting dependency information, less training data are required for training the model. We conducted a series of experiments by gradually reducing the scale of the training data. By contrast, the scale of the testing data remained unchanged. For comparison, the same testing data was used in the experiments.

The scales of the training data were marked as $s_i \in S$, and $S = \{100\%, 90\%, \ldots 20\%, 10\%\}$, indicating that s_i of the whole training data was used. For comparison, we take the BGRU-Att model, which is an evolved version of

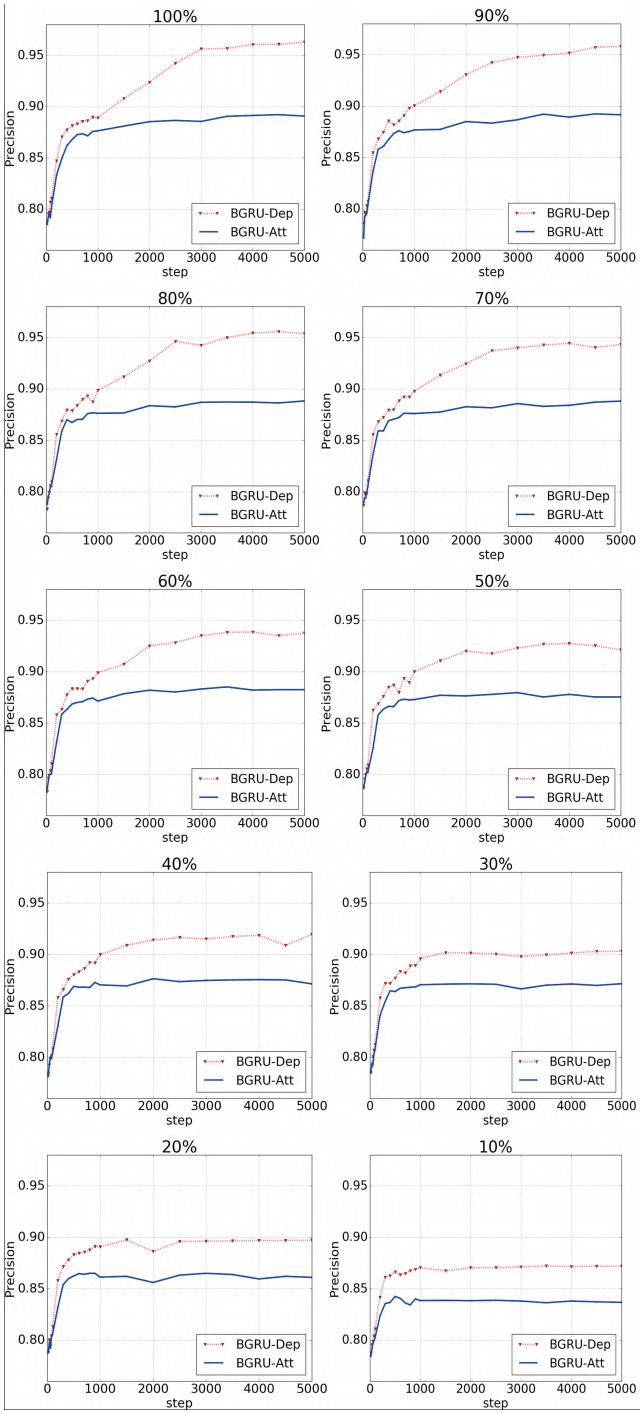


FIGURE 3. Precision over steps of each s_i in $S = \{100\%, 90\%, \ldots 20\%, 10\%\}$. The performance (0.896) of $s_i = 20\%$ in BGRU-Dep is higher than $s_i = 100\%$ in BGRU-Att (0.891).

(Lin *et al.* 2016). From the results shown in Figure 3, we can conclude the following. (1) For all $s_i \in S$, our model achieves a higher precision rate than BGRU-Att, proving the effect of dependency information. (2) Less data are required to obtain the same or even better result compared to methods without dependency information. For instance, the performance (0.896) of $s_i = 20\%$ in BGRU-Dep is higher than $s_i = 100\%$

TABLE 2. Result of ontology knowledge completion.

	Class	Axiom	Logical axiom	Annotation axiom	SubClassOf	EquivalentClasses	Object property	Hidden GCI
DO components relevant to gastroenterology	341	5458	407	4643	346	61	15	7
IASO ontology	398	5971	773	4731	403	199	15	32

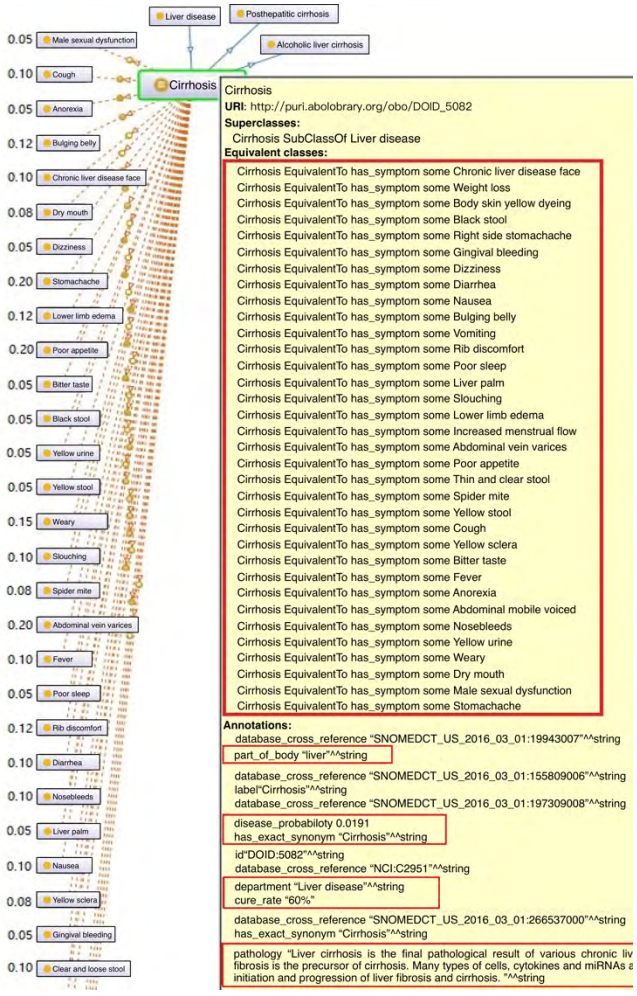


FIGURE 4. Ontology class: Hepatocirrhosis.

in BGRU-Att (0.891). Thus, with only one fifth of the training data, our model can achieve performance equivalent to that of BGRU-Att with the complete training data.

These results indicate that dependency features can provide abstract-level features, which can alleviate the requirement for a large amount of training data.

5) ONTOLOGY KNOWLEDGE COMPREHENSIVENESS

Ontology knowledge completion was carried out based on the initial DO gastroenterology ontology. This research studied 319 gastrointestinal diseases and 77 relevant symptoms. Seven new properties, namely body_part, probability, disease_probability, cure_rate, department, pathology and susceptible, and their corresponding relationships were added to the gastroenterology disease ontology. To improve the accuracy of reasoning, 31 prior possibilities of diseases and 171 conditional probabilities between diseases and syndromes were added to the ontology. Table 2 indicates the comparison of the ontology knowledge comprehensiveness between DO (gastroenterology disease relevant components) and the proposed IASO ontology.

The English names in Figure 4 are translations of their Chinese counterparts to facilitate comprehension. Take the disease 肝硬变 (cirrhosis) as an example. The solid lines represent the upper and lower levels of the class relationships, whereas the dotted lines show the relationships between diseases and symptoms.

The newly added information is highlighted by red borders. A number of new entities and relations are added to the ontology. Some examples of omissions from the DO ontology for the disease 'hepatocirrhosis' include the symptoms 'reddened palms' and 'lower extremity edema', which are considered highly relevant in the synonym identification. Similarly, the relation between the disease 'hepatocirrhosis' and lesions 'liver' are not listed in the DO ontology. This relation was suggested by our relation extraction algorithms, illustrating the potential for a data-driven approach to uncover unknown relations. The ontology knowledge completion results indicate that it is possible to construct a high-quality health ontology directly from text using synonym identification and relation extraction.

V. CONCLUSION

Ontology construction is to extract meaningful information from data sources. The results obtained in the previous section highlight the necessity of using synonym identification and relation extraction for clinical ontology construction.

Several points warrant further study in the next phase of research. First, experiments will be focused on more specific fields such as disease names, drugs, side effects, and complications to further improve the accuracy of word similarity measures. Second, we will improve relation extraction by considering the application of linguistic knowledge, such as inversion transduction grammar, tree substitution grammar and tree adjoining grammar.

REFERENCES

[1] R. Stevens, C. A. Goble, and S. Bechhofer, "Ontology-based knowledge representation for bioinformatics," *Briefings Bioinf.*, vol. 1, no. 4, pp. 398–414, 2000.

[2] M. A. G. Hazber, R. Li, X. Gu, G. Xu, and Y. Li, "Semantic SPARQL query in a relational database based on ontology construction," in *Proc. 11th Int. Conf. Semantics, Knowl. Grids (SKG)*, Aug. 2015, pp. 25–32.

[3] A. Arora, M. Singh, and N. Chauhan, "Automatic ontology construction using conceptualization and semantic roles," *Int. J. Inf. Retrieval Res.*, vol. 7, no. 3, pp. 62–80, 2017.

[4] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Vancouver, BC, Canada, 2008, pp. 1247–1250.

[5] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: A probabilistic taxonomy for text understanding," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Scottsdale, AZ, USA, 2012, pp. 481–492.

[6] Y. Xia, T. Zhao, J. Yao, and P. Jin, "Measuring Chinese-English cross-lingual word similarity with *HowNet* and parallel corpus," in *Proc. Int. Conf. Intell. Text Process. Comput. Linguistics*. Berlin, Germany: Springer, 2011, pp. 221–233.

[7] X.-W. Chen and X. Lin, "Big data deep learning: Challenges and perspectives," *IEEE Access*, vol. 2, pp. 514–525, 2014.

[8] H. Alani, "Position paper: Ontology construction from Online ontologies," in *Proc. 15th Int. Conf. World Wide Web*, 2006, pp. 491–495.

[9] J. Li, Z. He, and Q. Zhu, "An entropy-based weighted concept lattice for merging multi-source geo-ontologies," *Entropy*, vol. 15, no. 6, pp. 2303–2318, 2013.

[10] C. P. Abinaya and V. P. Sumathi, "Semi-automatic ontology merging of domain specific ontologies," *Int. J. Sci. Res.*, vol. 4, no. 1, pp. 1010–1012, 2013.

[11] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proc. 14th Conf. Comput. Linguistics*, Nantes, France, 1992, pp. 539–545.

[12] T. Weale, C. Brew, and E. Fosler-Lussier, "Using the wiktionary graph structure for synonym detection," in *Proc. Workshop People's Web Meets NLP: Collaboratively Constructed Semantic Resour.*, Suntec, Singapore, 2009, pp. 28–31.

[13] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Measuring semantic similarity between words using Web search engines," in *Proc. WWW*, 2007, pp. 757–766.

[14] J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana, and A. de Mendonça "Data mining methods in the prediction of dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests," *BMC Res. Notes*, vol. 4, no. 1, p. 299, 2011.

[15] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *Proc. AAAI*, vol. 1, 2006, pp. 775–780.

[16] M. Hagiwara, Y. Ogawa, and K. Toyama, "Supervised synonym acquisition using distributional features and syntactic patterns," *Inf. Media Technol.*, vol. 4, no. 2, pp. 558–582, 2009.

[17] A. Yates and O. Etzioni, "Unsupervised methods for determining object and relation synonyms on the Web," *J. Artif. Intell. Res.*, vol. 34, no. 1, pp. 255–296, 2009.

[18] R. Schwartz, R. Reichart, and A. Rappoport, "Symmetric pattern based word embeddings for improved word similarity prediction," in *Proc. 19th Conf. Comput. Natural Lang. Learn.*, 2015, pp. 258–267.

[19] C. N. dos Santos, B. Xiang, and B. Zhou. (May 2015). "Classifying relations by ranking with convolutional neural networks." [Online]. Available: https://arxiv.org/abs/1504.06580

[20] M. Surdeanu, J. Tibshirani, R. Nallapati, and C. D. Manning, "Multi-instance multi-label learning for relation extraction," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn.*, 2012, pp. 455–465.

[21] D. Zeng, K. Liu, Y. Chen, and J. Zhao, "Distant supervision for relation extraction via piecewise convolutional neural networks," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1753–1762.

[22] R. Socher, B. Huval, C. D. Manning, and A. Y. Ng, "Semantic compositionality through recursive matrix-vector spaces," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn.*, 2012, pp. 1201–1211.

[23] Y. Liu, F. Wei, S. Li, H. Ji, M. Zhou, and H. Wang. (Jul. 2015). "A dependency-based neural network for relation classification." [Online]. Available: https://arxiv.org/abs/1507.04646

[24] F. Cheng and Y. Miyao, "Classifying temporal relations by bidirectional lstm over dependency paths," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 1–6.

[25] G. Xun, Y. Li, J. Gao, and A. Zhang, "Collaboratively improving topic discovery and word embeddings by coordinating global and local contexts," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 535–543.

[26] K. Lei, S. Si, D. Wen, and Y. Shen, "An enhanced computational feature selection method for medical synonym identification via bilingualism and multi-corpus training," in *Proc. IEEE 2nd Int. Conf. Big Data Anal. (ICBDA)*, Mar. 2017, pp. 909–914.

[27] Y. Li, W. Li, F. Sun, and S. Li. (2015). "Component-enhanced Chinese character embeddings." [Online]. Available: https://arxiv.org/abs/1508.06669

[28] R. L. Cilibrasi and P. M. B. Vitanyi, "The Google similarity distance," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 370–383, Mar. 2007.

[29] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," *Comput. Sci.*, vol. 5, no. 1, p. 36, 2015.

[30] Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun, "Neural relation extraction with selective attention over instances," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 2124–2133.

[31] E. Terra and C. L. A. Clarke, "Frequency estimates for statistical word similarity measures," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Human Lang. Technol.*, vol. 1, 2003, pp. 165–172.

[32] P. D. Turney, "Mining the Web for synonyms: PMI-IR versus LSA on TOEFL," in *Proc. Eur. Conf. Mach. Learn.* Springer, 2001, pp. 491–502.

[33] K. Marek *et al.*, "The parkinson progression marker initiative (PPMI)," *Prog. Neurobiol.*, vol. 95, no. 4, pp. 629–635, 2011.

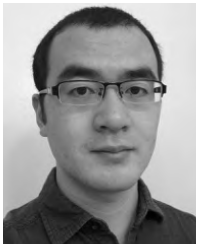
[34] I. Iacobacci, M. T. Pilehvar, and R. Navigli, "Sensembed: Learning sense embeddings for word and relational similarity," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics and 7th Int. Joint Conf. Natural Lang. Process.*, vol. 1, 2015, pp. 95–105.

[35] A. Handler, "An empirical study of semantic similarity in WordNet and Word2Vec," Ph.D. dissertation, Dept. Comput. Sci., Univ. New Orleans, New Orleans, LA, USA, 2014.

[36] K. Hu *et al.*, "A domain keyword analysis approach extending term frequency-keyword active index with Google Word2Vec model," *Scientometrics*, vol. 114, no. 3, pp. 1031–1068, 2018.

[37] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, "Relation classification via convolutional deep neural network," in *Proc. COLING*, 2014, pp. 2335–2344.

YING SHEN received the Erasmus Mundus master's degree in natural language processing from the University of Franche-Comté, France, and the University of Wolverhampton, U.K., and the Ph.D. degree specialized in medical and biomedical information science from the University of Paris Ouest Nanterre La Défense, France. She is currently an Assistant Professor with the School of Electronics and Computer Engineering, Peking University, where she leads the Artificial Intelligence–Knowledge Graph Team for medical research. Her research interests are mainly focused in the areas of medical informatics, natural language processing and machine learning.



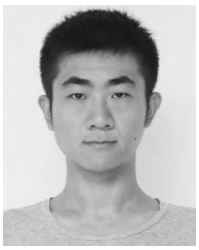
YALIANG LI received the Ph.D. degree in computer science from University at Buffalo, NY, USA, in 2017. He is broadly interested in machine learning, data mining and information analysis. In particular, he is interested in analyzing information from multiple heterogeneous sources, including but not limited to information integration, knowledge graph, anomaly detection, data stream mining, trustworthiness analysis, and transfer learning.

YANG DENG received the B.S. degree in computer science from the Beijing University of Posts and Telecommunications, in 2016. He is currently pursuing the M.S. degree in computer science with Peking University. His research interest is mainly focused in the area of deep learning and question answering.

JIN ZHANG received the B.S. degree in information security from the Wuhan University of China, in 2017. She is currently pursuing the M.S. degree in computer science with Peking University. Her research interest is mainly focused in the area of NLP and AI.

MIN YANG received the B.S. degree from Sichuan University in 2012 and the Ph.D. degree from The University of Hong Kong in 2017. She is currently an Assistant Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Science. Her current research interests include machine learning, deep learning, and natural language processing.

JINSONG CHEN received the Ph.D. degree from the University Institute of Lisbon, Portugal. He has been involved in hospital informationization for over 20 years. He was with the Peking University Shenzhen Hospital and the Hong Kong University Shenzhen Hospital. He is currently a Senior Engineer with the Shenzhen Maternal and Child Healthcare Hospital. His current research interests include medical decision making, medical informatics, and smart medicine.



SHANGCHUN SI received the B.S. degree in computer science from Peking University, China, in 2014, where he is currently pursuing the M.S. degree in computer science. His research interests mainly focus on construction and completion of medical knowledge graphs.

KAI LEI received the B.Sc. degree from Peking University, China, in 1998, the M.Sc. degree from Columbia University in 1999, and the Ph.D. degree from Peking University in 2015, all in computer science. He has worked for companies, including the IBM Thomas J. Watson Research Center, Citigroup, Oracle, and Google from 1999 to 2004. He currently is an Associate Professor with the School of Electronic and Computer Engineering, Peking University, Shenzhen. He has been participating in the CENI Project supported by the National Development and Reform Commission since 2016. His research interests include social networks, knowledge graphs, big data technologies, and named data networking.