IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 6, 2019, accepted December 1, 2019, date of publication December 5, 2019,
date of current version December 18, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2957827

Detecting Hypernymy Relations Between Medical Compound Entities Using a Hybrid-Attention Based Bi-GRU-CapsNet Model

CHENMING XU[1,3], **YANGMING ZHOU**[1,2], **QI WANG**[2], **ZHIYUAN MA**[2], **AND YAN ZHU**[3]

[1]Key Laboratory of Advanced Control and Optimization for Chemical Processes, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China
[2]School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China
[3]School of Science, East China University of Science and Technology, Shanghai 200237, China

Corresponding authors: Yangming Zhou (ymzhou@ecust.edu.cn) and Yan Zhu (zhuygraph@ecust.edu.cn)

This work was supported in part by the National Key Research and Development Program of China for Precision Medical Research under Grant 2018YFC0910500, in part by the Shanghai Sailing Program under Grant 19YF1412400, in part by the Fundamental Research Funds for the Central Universities under Grant 222201817006, and in part by the National Major Scientific and Technological Special Project for Significant New Drugs Development under Grant 2018ZX09201008.

ABSTRACT Named entities composed of multiple continuous words frequently occur in domain-specific knowledge graphs. In general, these named entities are composable and extensible, such as names of symptoms and diseases in the medical domain. Unlike the general entities, we address them as *compound entities*, and try to identify hypernymy relations between them. Hypernymy detection between compound entities plays a critical role in domain-specific knowledge graph construction. Traditional hypernymy detection approaches do not perform well on compound entities for two reasons. One is the lack of contextual information, and the other is the absence of compound entities, i.e. out-of-vocabulary (OOV) problem. In this paper, we propose a hybrid-attention-based method called Bi-GRU-CapsNet for the detection of hypernymy relations. The hybrid attention mechanism consists of heuristic attention and self-adaptive attention, which are used for the lack of contextual information. The attentions focus on the differences of two compound entities on the lexical and semantic level, respectively. For OOV problem, the English words or Chinese characters in compound entities are fed into bidirectional gated recurrent units (Bi-GRUs). Additionally, we use capsule network (CapsNet) to determine the existence of hypernymy relations under different cases. Experimental results show that our proposed method outperforms other baseline methods on both English and Chinese corpora of symptom and disease pairs.

INDEX TERMS Capsule network, medical compound entities, electronic health records, hybrid attention mechanism, hypernymy detection.

I. INTRODUCTION

Hnypernymy represents an important semantic relation, which is the backbone of almost every taxonomy, ontology and semantic network. As a result, hypernymy detection has many applications, such as taxonomy creation [1], ontology extension [2], question answering [3], machine reading [4], sentence similarity estimation [5] and text generation [6]. However, existing methods for hypernymy detection mainly handle the cases of one-word entities. While in domain-specific named entities consisted of multiple continuous words commonly occurred. In medical domain, the symptom and disease entities are composable and extensible. For example, "*carcinoma*" is a disease entity, and words describing body parts can be added to it, e.g. "*carcinoma of endocrine gland*". Also, the entity can be extended with descriptions, e.g. "*carcinoma of multiple endocrine glands*". If causes are included, the entity becomes "*carcinoma in situ of multiple endocrine glands*". Note that not all multiple-word entities are composable. For distinguishing purpose, we use *compound entities* for those that are composable.

Given a pair of entities (X_1, X_2), hypernymy detection aims to determine whether X_1 is a broad category that contains X_2. If the relation holds between entities X_1 and X_2,

The associate editor coordinating the review of this manuscript and approving it for publication was Mansoor Ahmed.

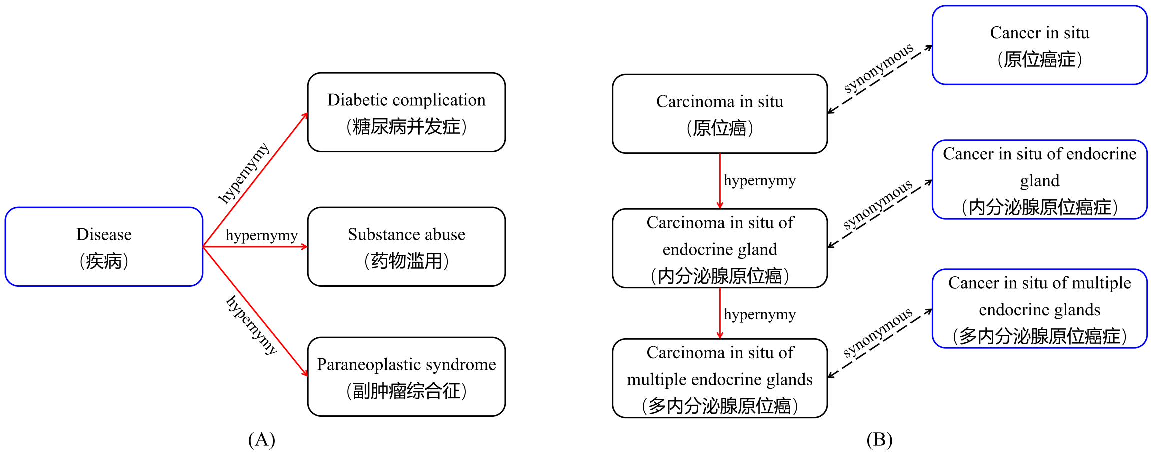


FIGURE 1. Examples to illustrate the hypernymy detection between general entities versus (see Subfigure (A)) compound entities (see Subfigure (B)), where the single arrow sign always points from a hypernym to a hyponym and the double arrow sign indicate that two compound entities are synonymous.

X_1 is a hypernym of X_2, and X_2 is a hyponym of X_1. Taken Figure 1 (a) for example, there are three hypernymy pairs including (*disease, diabetic complication*), (*disease, substance abuse*) and (*disease, paraneoplastic syndrome*). In contrast to hypernymy detection between general entities, this work is devoted to detecting hypernymy relations between compound entities. Taken Figure 1 (b) for example, there are multiple hypernymy pairs. Specifically, "*carcinoma in situ*" is a hypernym of "*carcinoma in situ of endocrine gland*", and "*carcinoma in situ of endocrine gland*" is also a hypernym of "*carcinoma in situ of multiple endocrine gland*".

Given different scenarios, there can be diverse categories of hypernymy pairs. In this paper, we roughly divide the hypernymy relations into four classes.

- **The hyponym is its hypernym with additional qualifiers**. For example, the hyponym "*carcinoma in situ of multiple endocrine gland*" has an additional qualifier "*multiple*" compared with its hypernym "*carcinoma in situ of endocrine gland*";
- **The headword of the hyponym is the hyponym of that of its hypernym**. For instance, in the hyponym "*carcinoma in situ of endocrine gland*" and the hypernym "*carcinoma of endocrine gland*", "*carcinoma in situ*" is the hyponym of "*carcinoma*";
- **The headword of the hyponym is the hyponym with additional qualifiers of that of its hypernym**. For example, in the hyponym "*carcinoma in situ of multiple endocrine glands*" and its hypernym "*carcinoma of endocrine gland*", "*carcinoma in situ*" is the hyponym of "*carcinoma*", and "*multiple*" is an additional qualifier;
- **The hyponym and the hypernym have synonyms**. For instance, in the hyponym "*cancer in situ*" and the hypernym "*carcinoma in situ of endocrine gland*", "*carcinoma*" and "*cancer*" are synonyms.

Detecting hypernymy relations from text is one of the important steps in the construction and enrichment of semantic resources. So far, a lot of methods have been proposed for this task. However, these methods are mainly for detecting hypernymy relations between general entities, which involve one-word entities, instead of compound entities. There are two main deficiencies when they are adapted for detecting hypernymy relations between compound entities. Firstly, they suffer from the out-of-vocabulary (OOV) problem because of the composability and extensibility of compound entities. Secondly, they do not take the different hypernymy relations into consideration.

To tackle the aforementioned problems, we propose a hybrid-attention based Bi-GRU-CapsNet, which is an extended version of our previous work [7]. Our proposed model has several important features. Firstly, we feed English words or Chinese characters in compound entities into bidirectional gated recurrent units (Bi-GRUs). Secondly, we use a hybrid-attention mechanism to focus on the dissimilar parts between the input pair. Thirdly, we apply capsule network (CapsNet), instead of softmax layer, to decide whether hypernym relation exists between medical compound entity pairs. On both Chinese and English corpora, we experimentally evaluate our proposed model. The main contributions of this paper can be summarized as follows.

- Unlike the general hypernymy detection, we define a new task for detecting hypernymy relations between compound entities, and develop a hybrid-attention based Bi-GRU-CapsNet model to solve it. Most existing methods try to use external knowledge to improve the performance of the model [9], [10], but compound entities are lack of external knowledge. Our proposed model utilizes the internal elements of compound entities, and it does not require any contextual information. A hybrid attention mechanism is proposed to identify the differences between two compound entities, which can effectively

improve the performance of our model. In addition, CapsNet is applied to hypernymy detection. Specifically, we employ capsules to deal with different cases of hypernymy relations between compound entities.
- Computational results show that our proposed model outperforms baseline methods both on Chinese and English corpora of symptom and disease pairs. Also, we build English and Chinese corpora of symptom and disease pairs for the task of detecting hypernymy relations between compound entities.

This paper is organized as follows. Section II outlines the related work in hypernymy detection. In Section III, our proposed method are introduced in detail. Experimental evaluation of the proposed method are provided in Section V, and Section VI concludes the paper and future directions.

II. RELATED WORK

Hypernymy detection has attracted considerable research effort in recent years, and several methods have been developed in the literature. There are two major approaches, namely pattern-based methods, distributional methods.

A. PATH-BASED METHODS

Path-based methods aim to identify hypernymy relations through the lexico-syntactic paths, which connect the joint occurrences of entity pairs in a large corpus [11]–[13]. Probably the first work is conducted by Hearst [11], who has found out that linking two noun phrases (NPs) via certain lexical constructions often implies hypernymy relations. For example, NP_1 is a hypernym of NP_2 in the lexical patterns ''NP_1 such as NP_2'' and ''NP_2 and other NP_1''. Some variations of pattern-based methods are also proposed to detect hypernymy relations [12], [14]. Recently, deep learning methods are applied in this task, in which the context paths are encoded by a recurrent neural network [13].

Path-based methods are simple and efficient. However, due to the ambiguity of a natural language and data sparsity, it is not robust to detect the hypernymy relations according to the context of entity pairs. Consider two sentences: ''various diseases could cause abnormalities of breathing, such as stridor, mouth breathing, and periodic breathing'' and ''various diseases could cause abnormalities of breathing, such as asthma, thalassaemia, and obstruction of respiratory tract''. According to the lexical patterns mentioned above, path-based methods can perform correctly for the former sentence but perform incorrectly for the latter, leading this kind of methods with low precision. Furthermore, people usually can not express every possible hypernymy relations in natural language texts, which leads to low precision and low recall rate. The path-based methods require co-occurrence of an entity pair, but there are many hypernymy pairs of compound entities in the same sentence. For Chinese symptoms, there are 324,253 sentences in the six Chinese healthcare websites mentioned in Section IV-A as well as the classification of disease symptoms in Baidu Baike.[1] However, among 12,800 Chinese symptom hypernymy pairs, there only exists 3,348 pairs co-occurring. The occurrence rate is merely 26.1%. Consequently, path-based methods are not applicable to hypernymy detection between compound entities.

B. DISTRIBUTIONAL METHODS

Distributional methods try to detect hypernymy relations based on the distributional representations of entity pairs, i.e. the contexts with which each entity occurs in the corpus.

Earlier distributional methods are usually based on **unsupervised learning**, which normally represent entities by their textual contexts in the form of sparse bag of words (SBOW) matrix, and employ a scoring function to detect hypernymy relations. Different scoring functions base on different hypotheses, e.g. distributional inclusion hypothesis [15], [16], distributional informativeness hypothesis [17] and selective distributional inclusion hypothesis [18]. An evaluation of scoring function can be found in [19]. Moreover, to solve the SBOW problem, Chang et al. compress the representations by matrix factorization [20].

The supervised methods can be categorized into classification methods, hypernym generation methods and ranking methods [21]. In **classification methods**, classifiers are employed to predict the relation of each pair, which is represented as embedding vectors by pre-trained neural language models [22], [23]. To aggregate the term pairs, the concat model [24], the diff model [25], the asym model [18] and the simDiff model [26] are proposed. Some works focus on using knowledge to improve the representations by considering the taxonomic structure of concepts [9], [27] or Hearst patterns [29]. **Projection learning methods** (a.k.a Hypernym generation methods) mainly learn a model which maps term embeddings to their hypernyms [30]–[32]. Wang et al. combine classification and hypernym generation methods by employing multiple fuzzy orthogonal projections and a neural network [33]. **Ranking methods** select the most probable hypernym for an entity, which performs poorly in recall [34]. However, some literatures show these methods work for Chinese language [35].

Compared with path-based methods, distributional methods normally perform better, because their performances do not rely on entity pairs' co-occurrence. However, distributional methods suffer from lexical memorization problem [36], which tends to recognize prototypical hypernyms as right answers. Meanwhile, a compound entity usually contains a group of words instead of a single word. When encoding compound entities into vectors through the information of their context, these methods often suffer from the out-of-vocabulary (OOV) problem for the compound entities maybe not in the corpus. Taken Chinese symptoms for example, among 26,821 Chinese symptoms, there exist 7,511 symptoms disappearing in the sentences. The miss rate is up to 28.0%. To overcome the prototypical hypernym

[1] https://baike.baidu.com/wikitag/taglist?tagId=75953/

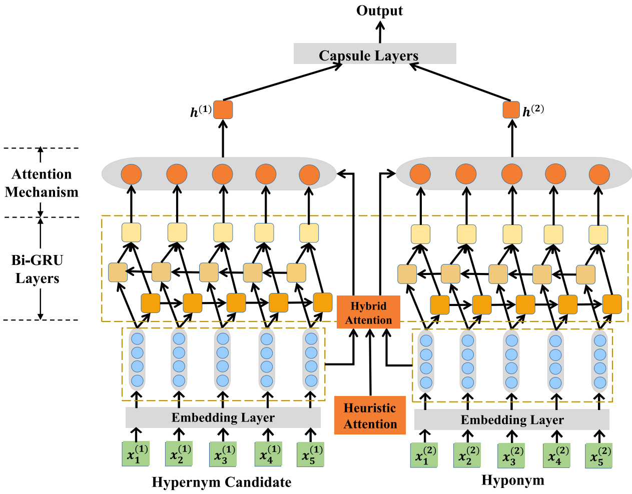


FIGURE 2. The architecture of the hybrid-attention based Bi-GRU-CapsNet model.

problem, we construct the negative samples with prototypical hypernyms(see Section IV-A). Moreover, most of the general-domain entities are one-word which causes the existing methods to focus on introducing external knowledge into the model. And there exist a large number of compound entities in the specific domain which is lacking external knowledge and has much structure and semantic information. To overcome the OOV problem and utilize the structure and semantic information of compound entities, we propose a hybrid attention mechanism based on the feature of medical compound entities (see Section III-D). Furthermore, we employ the capsule network in the model, which can not only represent the probability, but also the cases of relations (see Section III-E).

III. HYBRID-ATTENTION BASED Bi-GRU-CapsNet MODEL

A. OVERVIEW

The architecture of our proposed model is given in Figure 2. In the model, we firstly send English words or Chinese characters into embedding layers to get their vector representations (see Section III-B), and then feed them into the bidirectional recurrent layers. For the recurrent layer, we employ gated recurrent units (GRUs) [37] (see Section III-C), because they have similar functionalities but fewer parameters, when compared with long short-term memory (LSTM) [8]. Afterwards, considering the similarity between entities we get the feature vector of each entity through a hybrid attention mechanism (see Section III-D). Finally, we use the output vectors of capsule network (CapsNet) (see Section III-E) to determine whether medical compound entity pairs have hypernymy relations. The key components are detailed in the following subsections.

B. EMBEDDING LAYER

Given a hypernym candidate X_1 with a sequence of T_1 words, and a hyponym X_2 with is a sequence of T_2 words,[2] the first step is to map discrete language symbols to distributed embedding vectors. Formally speaking, for each word $x_t^{(k)}$, we obtain corresponding embedding vector $\boldsymbol{e}_t^{(k)} \in \mathbb{R}^{d_e}$ from embedding matrix, where $k \in \{1, 2\}$ and, $k = 1$ or $k = 2$ respectively indicates that the sequence is a hypernym candidate or a hyponym, $t \in \{1, 2, \ldots, T_k\}$ identifies the position of $x_t^{(k)}$ in X_k, and d_e is a hyper-parameter indicating the size of word embedding.

[2] X_1 and X_2 can either consist of English or Chinese. Normally, they are mentioned as sequences of words.

C. Bi-GRU LAYER

The gated recurrent unit (GRU) is an important variant of recurrent neural network, and was originally proposed by Cho *et al.* [37]. Given an input $\boldsymbol{x}_t$ and the previous state $\boldsymbol{h}_{t-1}$ at position t, $\boldsymbol{h}_t$ can be computed as follows.

$$\boldsymbol{r}_t = \sigma(\boldsymbol{W}_r\boldsymbol{x}_t + \boldsymbol{U}_r\boldsymbol{h}_{t-1}) \tag{1}$$
$$\boldsymbol{u}_t = \sigma(\boldsymbol{W}_u\boldsymbol{x}_t + \boldsymbol{U}_u\boldsymbol{h}_{t-1}) \tag{2}$$
$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W}_c\boldsymbol{x}_t + \boldsymbol{U}(\boldsymbol{r}_t \odot \boldsymbol{h}t - 1)) \tag{3}$$
$$\boldsymbol{h}_t = (1 - \boldsymbol{u}_t) \odot \boldsymbol{h}_{t-1} + \boldsymbol{u}_t \odot \tilde{\boldsymbol{h}}_t \tag{4}$$

where $\boldsymbol{h}_t$, $\boldsymbol{r}_t$ and $\boldsymbol{u}_t \in \mathbb{R}^d$ are d-dimensional hidden state, reset gate, and update gate, respectively; $\boldsymbol{W}_r$, $\boldsymbol{W}_u$, $\boldsymbol{W}_c \in \mathbb{R}^{d_e \times d}$ and $\boldsymbol{U}_r$, $\boldsymbol{U}_u$, $\boldsymbol{U} \in \mathbb{R}^{d \times d}$ are the parameters of the GRU; σ is the sigmoid function, and $\odot$ indicates element-wise production.

For the t-th word in a sequence, we use hidden states $\overrightarrow{\boldsymbol{h}_t}$ and $\overleftarrow{\boldsymbol{h}_t}$, which are encoded through the forward GRU and the backward GRU, to represent the preceding and following context of x_t, respectively. The concatenation $\boldsymbol{h}_t = [\overrightarrow{\boldsymbol{h}_t}; \overleftarrow{\boldsymbol{h}_t}]$ is the output of the Bi-GRU layer at t.

D. ATTENTION MECHANISM

For hypernymy detection, it is useful to focus on only the different parts between two compound entities. To achieve it, we propose a heuristic attention mechanism to mask the longest common sequence between two entities. A soft way to focus on the dissimilar parts between two entities is to calculate the similarity score of their embeddings. Moreover, these two attention mechanisms can be combined together such that the advantages of heuristic rule and self-adaptive method can be fully utilized. In our model, we propose a hybrid attention mechanism to improve detection performance. The feature vector $\boldsymbol{h}^{(k)}$ of the entity X_k is defined as a weighted sum, which is computed as follows:

$$\boldsymbol{h}^{(k)} = \sum_{i=1}^{T_k} \alpha_i^{(k)} \boldsymbol{h}_i^{(k)} \tag{5}$$

where $\boldsymbol{h}_i^{(k)}$ is the output of the Bi-GRU layer at i, and the weight $\alpha_i^{(k)}$ is computed according to following three different attention mechanisms.

1) HEURISTIC ATTENTION

$$\alpha_i^{(k)} = aw_i^{(k)} + b \tag{6}$$

$$w_i^{(k)} = \begin{cases} 0, & x_i^{(k)} \in seqLCS \\ 1, & x_i^{(k)} \notin seqLCS \end{cases} \tag{7}$$

seqLCS refers to the longest common subsequence of X_1 and X_2, $x_i^{(k)}$ is the i-th word of X_k, and a, b are the network parameters. The heuristic attention has a great effect on the entity pairs when X_1 is similar with X_2. However, it simply focuses on the different lexical parts between X_1 and X_2, rather than on the semantic aspect.

2) SELF-ADAPTIVE ATTENTION

$$\alpha_i^{(k)} = 1 - \max_j\{similarity(e_i^{(k)}, e_j^{(z(k))^T})\} \tag{8}$$

Here, $k \in \{1, 2\}$ and $z(k) \in \{1, 2\}\backslash\{k\}$, $i \in \{1, 2, \ldots, T_k\}$, $j \in \{1, 2, \ldots T_{z(k)}\}$. We firstly calculate a similarity between the hypernym candidate and the hyponym, then make it minus 1, to focus on only the different parts.

3) HYBRID ATTENTION

$$\alpha_{Heuristic}^{(k)} = (\alpha_{1,1}^{(k)}, \alpha_{2,1}^{(k)}, \ldots, \alpha_{T_k,1}^{(k)}) \tag{9}$$

$$\alpha_{Self-adaptive}^{(k)} = (\alpha_{1,2}^{(k)}, \alpha_{2,2}^{(k)}, \ldots, \alpha_{T_k,2}^{(k)}) \tag{10}$$

$$\alpha_i^{(k)} = \alpha_{Self-adaptive}^{(k)} \odot \alpha_{Heuristic}^{(k)} \tag{11}$$

$\odot$ is the symbol of element-wise product. Our hybrid attention is obtained by hybridizing heuristic attention and self-adaptive attention. Specifically, we keep the value of self-adaptive attention still in the different parts and let it be zero in the same parts.

As mentioned earlier, the differences between X_1 and X_2 help us to detect hypernymy. Thus, it is better to assign larger attention weights for the dissimilar parts while smaller weights for similar ones (see Equation (5)). To achieve the goal, we tried several methods of assigning attention, eventually, we found that the combination of heuristic method and self-adaptive method had the best result (see Table 5). The attention mechanism can also improve unseen entity pairs problems because of the use of semantic similarity.

E. CAPSULE LAYER

The capsule layer was first proposed in [38] for digit recognition. As illustrated in Figure 3, a non-linear squashing function is used to ensure that short vectors are shrunk to almost zero and long vectors are shrunk to a value slightly below 1.

$$\boldsymbol{v}_j = \frac{||\boldsymbol{s}_j||^2}{1 + ||\boldsymbol{s}_j||^2} \frac{\boldsymbol{s}_j}{||\boldsymbol{s}_j||} \tag{12}$$

where $\boldsymbol{v}_j$ is the output of capsule $\boldsymbol{s}_j$.

$$\boldsymbol{s}_j = \sum_{i=1} c_{ij} \widehat{\boldsymbol{u}}_{j|i} \tag{13}$$

$$\widehat{\boldsymbol{u}}_{j|i} = \boldsymbol{W}_{ij} \boldsymbol{u}_i \tag{14}$$

where c_{ij} is coupling coefficient, and it is determined by an iterative dynamic routing algorithm with a hyper-parameter, i.e. the number of iterations r.

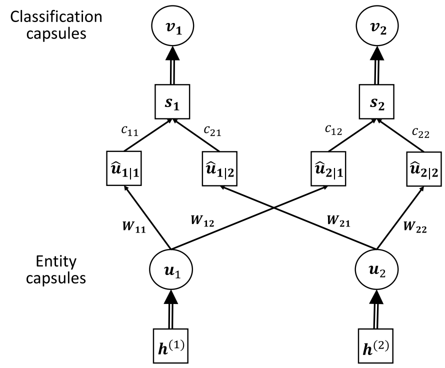


FIGURE 3. The structure of capsule layers.

TABLE 1. Details of english and chinese corpora.

Corpus		Positive	Negative	Total
English	Training set	27,872	27,872	55,744
	Test set	9,954	9,954	19,908
	Validation set	1,991	1,991	3,982
Chinese	Train set	8,960	8,960	17,920
	Test set	3,200	3,200	6,400
	Validation set	640	640	1,280

We use the margin loss L_j as the loss function. For each classification capsule $\boldsymbol{v}_j$, the network tries to minimize a L_j in training phase:

$$L_j = R_j \max(0, m^+ - ||\boldsymbol{v}_j||)^2 + (1 - R_j) \max(0, ||\boldsymbol{v}_j|| - m^-)^2 \tag{15}$$

where $j = 0$ means there is a hypernymy relation between two compound entities, otherwise $j = 2$. $R_j = 1$ iff the corresponding relation j exists and $m^+ = 0.9$ and $m^- = 0.1$. The total loss is simply the sum of the losses of both classification capsules.

IV. COMPUTATIONAL STUDIES

A. DATASETS

We evaluate our proposed model on both English and Chinese corpora .[3] The corpora consist of symptom and disease compound entity pairs, and the statistical information is presented in Table 1.

The English corpus is obtained by extracting clinical finding pairs in SNOMED CT [39]. SNOMED CT is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world. It is easy to get hypernymy compound entities for its hierarchical relation structure. We construct the entity pairs with two neighbour layers. Clinical finding pairs with hypernymy relations are chosen as positive instances. The negative instances consist of hyponymy and unrelated pairs. The hyponymy pairs are the symmetric forms of the hypernymy pairs. The unrelated pairs

[3]https://github.com/ECUST-NLP-Lab/medicalHypernymy

include two unrelated symptom entities. To solve the prototypical hypernym problem, we select the clinical findings that are always hypernyms or hyponyms, and use them as hyponymy pairs and unrelated pairs. To reduce the influence of heuristic attention, we use clinical findings that are similar to construct unrelated pairs.

The Chinese version of SNOMED CT is not available, so we build one from six main Chinese healthcare websites, by extracting hypernymy and synonymy pairs from Chinese medical text and semi-structured Chinese medical data. Hypernymy pairs are regarded as positive instances, hyponymy, synonymy and unrelated pairs are considered as negative instances. The hyponymy pairs and unrelated pairs are constructed in the same way as English corpus does. The synonymy pairs are constructed based on a self-defined symptom synonymy dictionary.

B. EXPERIMENTAL SETTINGS

Due to the lack of context, we randomly initialize word embeddings into 256-dimensional vectors. The size of GRU hidden states and the dimension of the capsules are both set to 64. The iteration number of routing between entity capsules and classification capsules is 2, and all the routing logics are initialized to zero. The model is trained by an adaptive method called AdaDelta [40] to minimize the margin loss, and the batch size is 128. Both the selection of margin loss function and the number of iterations $r = 2$ is determined according to experimental results presented in Section V-A and Section V-B. The hybrid attention mechanism combines heuristic attention and adaptive attention. For the heuristic attention, we set $a = 1$ and $b = 0$. We let each medical compound entity end with a special end-of-term symbol "$\langle EOS\rangle$", and the attention weight $w_i^{(k)}$ of "$\langle EOS\rangle$" is set to 1, which enables the model to have a non-zero output vector of an entity. We compare four attention mechanisms in Section V-C. The best results in the tables are in bold in the following sections. Several widely-used measures, namely Precision, Recall, and F_1-score, are employed to evaluate the algorithms in the following experiments [41], [42].

C. COMPARISON WITH BASELINE APPROACHES

We experimentally compare our proposed model with several baseline methods. These baseline algorithms are selected based on the following reasons. Consider the first case of hypernymy relations, we compare our proposed model with two basic methods, namely string containing method and set containing method. Due to the lack of context corpus of medical compound entity pairs, existing methods cannot be directly applied to this task, such as [9], [10]. As an adaptation, we select several state-of-the-art supervised distributional approaches and distributed approaches as reference algorithms. Specifically, as to supervised distributional approaches, instead of training entity embeddings on context corpus via word2vec [22], we randomly initialize *whole entity* embeddings, and update them through the supervised learning process. In addition, we also use the *sum of word* embeddings as the entity vector for solving the OOV problem. These baseline methods can be summarized as following two categories.

1) **Basic methods:**
   - **String containing method**: given a hypernym candidate X_1 and a hyponym X_2, if X_2 contains X_1, X_1 and X_2 are considered to have hypernymy relations.
   - **Set containing method**: given a hypernym candidate X_1 and a hyponym X_2, if the English word (or Chinese character) set of X_2 contains the English word (or Chinese character) set of X_1, X_1 and X_2 are considered to have hypernymy relations.
2) **Distributional methods:**
   - **Feature vector method** [18], [24], [25] is a distributional method, where an entity pair is represented by a feature vector. The feature vector is a combination of entity embeddings. Following Shwartz *et al.* [13], we test three state-of-the-art combination methods, namely concatenation (i.e., $\overrightarrow{X_1} \oplus \overrightarrow{X_2}$), difference (i.e., $\overrightarrow{X_1} - \overrightarrow{X_2}$) and dot-product (i.e., $\overrightarrow{X_1} \cdot \overrightarrow{X_2}$). For each kind of feature vectors, we train a number of classifiers: logistic regression, support vector machine (SVM), and SVM with radial basis function (RBF) kernel. We perform model selection on the validation set to select the best one.
   - **Projection learning method** [25] is also a distributional method which trains projection matrices between entity embeddings to predict hypernymy.[4]
   - **Simple RNN method** [43] is a distributional method which passes entity pairs to an RNN to predict hypernymy.
   - **Term embedding method** [28] is a supervised distributional method with negative sampling to learn entity embeddings via pre-extracted hypernymy pairs.

Table 2 presents detailed comparative results of our proposed model and baseline methods. From this table, we find that our proposed model outperforms these baseline algorithms. Specifically, our proposed model achieves the best F_1-scores on both English and Chinese corpora compared to all baseline algorithms. In terms of Recall and Precision, our proposed model also achieves competitive performance compared to baseline algorithms.

[4]They clustered the embedding space with k-means for domain adaptation, tuning k based on a validation set. However, the clustering process relying on entity embeddings is not applicable for our task due to the lack of context information. That is, entity embeddings are produced via word2vec which needs context information as to its input. Fortunately, Fu *et al.* [25] stated that different clusters correspond to different kinds of entities, such as animals and people's occupations. Since our corpora contain only one kind of entities, i.e. clinical findings, it is not necessary to cluster the embedding space.

TABLE 2. Comparative results of our proposed model and baseline methods on both english and chinese corpora.

Methods		English corpus			Chinese corpus		
		Precision	Recall	F_1-score	Precision	Recall	F_1-score
String containing		**95.20**	2.39	4.66	**99.17**	78.47	87.61
Set containing		94.45	5.64	10.66	97.95	86.66	91.96
Feature vector*	Whole entity	73.72	81.77	77.54	79.51	69.47	74.15
	Sum of words	87.31	88.95	88.12	93.35	**94.78**	94.06
Projection learning	Whole entity	70.96	83.47	76.71	78.19	64.75	70.84
	Sum of words	83.30	87.22	85.21	85.32	86.84	86.08
Simple RNN	Whole entity	76.51	80.78	78.59	81.42	62.59	70.78
	Sum of words	88.52	89.92	89.22	80.11	63.56	70.88
Term embedding	Whole entity	29.81	59.62	39.75	78.19	64.75	70.84
	Sum of words	42.63	85.25	56.83	38.78	77.56	51.73
Hybrid-attention based Bi-GRU-CapsNet		90.81	**92.21**	**91.51**	97.37	93.62	**95.46**

* We only give the results of feature vector methods with the best settings (i.e., combination method is the difference ($\overrightarrow{X_1} - \overrightarrow{X_2}$) and classifier is SVM with RBF kernel).

For two basic methods, we observe that they work well on medical compound entity pairs in Chinese corpus, but perform badly on English corpus, with the results of Recall ($\leqslant$ 5.64%) and F_1-score ($\leqslant$ 10.66%). It seems reasonable that most of the Chinese symptom hypernymy pairs belong to the first case, i.e. qualifiers added, while English corpus does not, so the performance in terms of recall declines sharply.

For four adapted baseline methods, we find that they achieve similar performance ($\geqslant$ 62.59%) on all three performance metrics except the term embedding method. In contrast to the whole entity version, all four adapted baseline algorithms which apply the sum vector of words, achieve better performance excluding one exception that term embedding method on Chinese corpus. In general, our proposed model roughly exceeds four adapted baseline algorithms except one. That is, the Recall of feature vector methods with the sum of word embeddings on Chinese corpus is 94.78%, which is slightly better than that of our proposed model (i.e., 93.62%).

Besides, we observe that simple RNN method with the sum of words performs better than the other baseline methods on English corpus, whose F_1-score is merely 2.29% lower than our proposed model because it utilizes an RNN to remember both the hyponym and its hypernym candidate when judgment. However, its F_1-score is far below our proposed model on the Chinese corpus. As mentioned above, most of the Chinese symptom hypernymy pairs belong to the first case, i.e. qualifiers added, and the recall of simple RNN method is nearly 20% lower than its Precision. It indicates that simple RNN method will be confused if the two entities are similar, and tends to think they have no hypernymy relation. While our proposed model employs a hybrid attention mechanism to focus on the different parts between entities, so it will not be confused even though the two entities are similar and is nearly 25% higher than simple RNN method in terms of F_1-score. Actually, except feature vector method with the sum of words, other baseline methods all perform poorly on the Chinese corpus. It is because only the feature vector method (i.e., difference + SVM with RBF kernel) pays special attention to the different parts between two entities. These interesting observations confirm the effectiveness of our proposed model for detecting hypernymy relations between compound entities.

TABLE 3. Comparisons between Bi-GRU models with the capsule layers, sigmoid layers and softmax layers under different loss functions on the English corpus.

Layer	r	Cross entropy Preci-sion	Cross entropy Recall	Cross entropy F_1-score	Margin loss Preci-sion	Margin loss Recall	Margin loss F_1-score
Sigmoid	-	89.00	90.45	89.72	88.90	89.65	89.28
Softmax	-	87.88	91.05	89.44	86.76	92.10	89.35
Capsule	1	89.50	90.69	90.09	89.50	90.71	90.01
	2	91.03	**91.25**	91.14	**90.81**	**92.21**	**91.51**
	3	**91.54**	91.04	**91.29**	90.30	91.07	90.68

V. DISCUSSION AND ANALYSIS

In this section, three groups of experiments are made to respectively investigate the effect of the capsule layers, interest of Bi-GRUs, and the effectiveness of the hybrid attention mechanism. Note that all the following experimental investigations are based on English corpus.

A. EFFECT OF THE CAPSULE LAYERS

To study the effect of the capsule layers integrating into our model, we experimentally compare the performance of Bi-GRU models with capsule layers, sigmoid layers,and softmax layers. We report the results on two different loss functions based on the English corpus. The detailed results are shown in the Table 3.

Table 3 shows the results of Bi-GRU models with the capsule layers, sigmoid layers and softmax layers under two different loss functions on the English corpus. From this table, we can obtain the following observations. The performance of Bi-GRU models with capsule layers is significantly better than the performance of Bi-GRU models with sigmoid layers and softmax layers on both two loss functions. It is reasonable that a capsule can learn a more robust representation for hypernymy detection, which can successfully deal with different cases in hypernymy relation. For the capsule

TABLE 4. Comparative results of models with or without Bi-GRUs under different loss functions on the English corpus.

Layer	Bi-GRUs	Cross entropy Preci-sion	Cross entropy Recall	Cross entropy F_1-score	Margin loss Prec-ision	Margin loss Recall	Margin loss F_1-score
Sigmoid	×	82.35	86.17	84.22	81.55	87.18	84.27
	✓	89.00	90.45	89.72	88.90	89.65	89.28
Softmax	×	82.49	85.93	84.17	81.62	86.75	84.11
	✓	87.88	91.05	89.44	86.76	92.10	89.35
Capsule*	×	85.96	90.67	88.25	87.28	87.43	87.36
	✓	**91.03**	**91.25**	**91.14**	**90.81**	**92.21**	**91.51**

* The result of the proposed model with capsule layers is obtained when the routing iteration is 2.

layers with different numbers of routing iterations (i.e., $r \in \{1, 2, 3\}$), we obverse that the best performance is achieved when the routing iteration is 2 (see the results in bold). For the results of capsule layers with different loss functions, we see that the results of margin loss are better than that of cross entropy. These observations confirm the interest of capsule layers. Consequently, we use the capsule layers and set the number of routing iterations $r = 2$ in our model.

B. INTEREST OF THE Bi-GRUs

To analyze the interest of the Bi-GRUs, we compare the computational results of sigmoid layers, softmax layers and capsule layers for the cases with or without Bi-GRUs. Based on the English corpus, we also discuss the results obtained by using two different loss functions. The computational results are shown in Table 4.

From Table 4, we observe that the use of Bi-GRUs significantly improves the detection performance. For example, in case of cross entropy, the performance of capsule layers without Bi-GRUs in terms of precision, recall and F_1-score is 85.96%, 90.67% and 88.25% respectively. With the help of Bi-GRUs, the performance of capsule layers in terms of precision, recall and F_1-score increases to 91.03%, 91.25% and 91.14% respectively. It is also true for the case with a margin loss function. For the sigmoid layers and softmax layers, the same observations can also be obtained. These observations prove the interest of Bi-GRUs to improve detection performance. In addition, we also obtain the same observation on the loss function as well as Section V-A. That is, the results of the margin loss function are better than that of cross entropy. Therefore, we use margin loss as our loss function in our model.

C. EFFECTIVENESS OF HYBRID ATTENTION MECHANISM

The attention mechanism is a very effective method for solving NLP tasks [44]. It was originally proposed by Bahdanau *et al.* [45] to deal with neural machine translation, where a softmax function is employed to calculate the attention weights. This section is devoted to investigating the effectiveness of our hybrid attention mechanism. Based on English corpus, we experimentally compare the following four attention mechanisms:

TABLE 5. Comparisons of four attention mechanisms on english corpus.

	Precision	Recall	F_1-score
Equal attention	85.71	89.72	87.67
Heuristic attention [7]	89.30	91.42	90.35
Self-adaptive attention	89.43	90.94	90.18
Hybrid attention	**90.81**	**92.21**	**91.51**

TABLE 6. Performance comparisons of four attention mechanisms on three general-domain corpora.

	BLESS Val	BLESS Test	EVALution Val	EVALution Test	ROOT9 Val	ROOT9 Test
Equal attention	**68.46**	**0.70**	69.91	**71.36**	79.61	80.44
Heuristic attention	**68.46**	**0.70**	69.65	68.77	79.61	80.44
Self-adaptive attention	**68.46**	0.62	66.67	59.95	72.42	71.54
Hybrid attention	**68.46**	0.62	**70.0**	70.08	**80.47**	**80.97**

1) Equal attention. It assigns the same attention to each word;
2) Heuristic attention. It compares two sequences and focuses on the different parts of them [7];
3) Self-adaptive attention. It uses a dot-product method to measure the similarity;
4) Hybrid attention. It hybridizes heuristic attention and self-adaptive attention.

Table 5 presents the detailed results of four attention schemes. From this table, we find that our model with hybrid attention mechanism achieves the best performance in all three measures. In addition, we observe that equal attention has the worst performence which means attention mechanism can effectively improve the performence of the model. The model with heuristic attention [7] achieves great progress compared to our model with self-adaptive attention except for Precision which confirm our heuristic attention mechanism is able to focus on the different parts between medical compound entity pairs. And the self-adaptive attention is a useful mechanism which improves the performence of the previous model. These interesting observations confirm the effectiveness of our hybrid attention mechanism.

To evaluate the impact of data, Table 6 shows the experimental results of four attention mechanisms on three general-domain corpora. The datasets used in the table are widely adopted in other literatures [46]–[48]. We split the datasets into 6:2:2. In validation sets, the hybrid mechanism achieves the highest f1-scores, while equal attention mechanism has the highest f1-scores in test sets. Note that our mechanism is designed based on the characteristics of compound entity pairs, and general-domain datasets are of one-word entity pairs. Therefore, the proposed mechanism does not significantly outperform other attentions. However, with proper division of datasets the proposed method can still achieve similar performance.

VI. CONCLUSIONS AND FUTURE WORK

Compound entities frequently occur in specific domains, but existing hypernymy detection methods do not perform well on them. In this paper, we propose a hybrid-attention based Bi-GRU-CapsNet model for hypernymy detection between medical compound entities. Our proposed model has three important features: 1). English words or Chinese characters in medical compound entities are separately encoded by Bi-GRUs to resolve the OOV problem; 2). A hybrid attention mechanism is proposed to only focus on the differences between two medical compound entities; 3). CapsNet is employed to handle different cases in hypernymy relation.

Extensive results show that our proposed model is able to achieve better performance compared to baseline methods. Computational results also verify the effectiveness of several important components integrated in our proposed model. For future work, one interesting direction is to improve synonymy inference for hypernymy detection.

REFERENCES

[1] R. Snow, D. Jurafsky, and A. Y. Ng, "Semantic taxonomy induction from heterogenous evidence," in *Proc. Int. Conf. Comput. Linguistics Annu. Meeting Assoc. Comput. Linguistics*, Sydney, NSW, Australia, Jul. 2006, pp. 801–808.

[2] J. Lee, J. Kim, and J. C. Park, "Automatic extension of gene ontology with flexible identification of candidate terms," *Bioinformatics*, vol. 22, no. 6, pp. 665–670, 2006.

[3] P. McNamee, R. Snow, P. Schone, and J. Mayfield, "Learning named entity hyponyms for question answering," in *Proc. Int. Joint Conf. Natural Language Process.*, Hyderabad, India, Jan. 2008, pp. 799–804.

[4] O. Etzioni, M. Banko, and M. J. Cafarella, "Machine reading," in *Proc. Nat. Conf. Artif. Intell. Innovative Appl. Artif. Intell. Conf.*, Boston, MA, USA, Jul. 2006, pp. 1517–1519.

[5] G. Soganciogl u, H. Öztürk, and A. Özgür, "BIOSSES: A semantic sentence similarity estimation system for the biomedical domain," *Bioinformatics*, vol. 33, no. 14, pp. i49–i58, 2017.

[6] O. Biran and K. McKeown, "Classifying taxonomic relations between pairs of Wikipedia articles," in *Proc. Int. Joint Conf. Natural Lang. Process.*, Nagoya, Japan, Oct. 2013, pp. 788–794.

[7] Q. Wang, C. Xu, Y. Zhou, T. Ruan, D. Gao, and P. He, "An attention-based bi-gru-capsnet model for hypernymy detection between compound entities," in *Proc. Int. Conf. Bioinf. Biomed.*, Madrid, Spain, Dec. 2018, pp. 1031–1035.

[8] Q. Wang, Y. Zhou, T. Ruan, D. Gao, Y. Xia, and P. He, "Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition," *J. Biomed. Informat.*, vol. 92, Apr. 2019, Art. no. 103133.

[9] T. L. Anh, Y. Tay, S. C. Hui, and S. Ki Ng, "Learning term embeddings for taxonomic relation identification using dynamic weighting neural network," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Austin, TX, USA, Nov. 2016, pp. 403–413.

[10] Y. Li, R. Zheng, T. Tian, Z. Hu, R. Iyer, and K. Sycara, "Joint embedding of hierarchical categories and entities for concept categorization and dataless classification," Jul. 2016, *arXiv:1607.07956*. [Online]. Available: https://arxiv.org/abs/1607.07956

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

[12] R. Snow, D. Jurafsky, and A. Y. Ng, "Learning syntactic patterns for automatic hypernym discovery," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1297–1304.

[13] V. Shwartz, Y. Goldberg, and I. Dagan, "Improving hypernymy detection with an integrated path-based and distributional method," Mar. 2016, *arXiv:1603.06076*. [Online]. Available: https://arxiv.org/abs/1603.06076

[14] A. Ritter, S. Soderland, and O. Etzioni, "What is this, anyway: Automatic hypernym discovery," in *Proc. AAAI Spring Symp., Learn. Reading Learn. Read*, Stanford, CA, USA, Mar. 2009, pp. 88–93.

[15] L. Kotlerman, I. Dagan, I. Szpektor, and M. Zhitomirsky-Geffet, "Directional distributional similarity for lexical inference," *J. Natural Lang. Eng.*, vol. 16, no. 4, pp. 359–389, Oct. 2010.

[16] A. Lenci and G. Benotto, "Identifying hypernyms in distributional semantic spaces," in *Proc. Joint Conf. Lexical Comput. Semantics*, Montreal, QC, Canada, Jun. 2012, pp. 75–79.

[17] E. Santus, A. Lenci, Q. Lu, and S. Schulte im Walde, "Chasing hypernyms in vector spaces with entropy," in *Proc. Conf. Eur. Chapter Assoc. Comput. Linguistics*, Gothenburg, Sweden, Apr. 2014, pp. 38–42.

[18] S. Roller, K. Erk, and G. Boleda, "Inclusive yet selective: Supervised distributional hypernymy detection," in *Proc. Int. Conf. Comput. Linguistics, Tech. Papers*, Dublin, Republic of Ireland, Aug. 2014, pp. 1025–1036.

[19] V. Shwartz, E. Santus, and D. Schlechtweg, "Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection," Dec. 2016, *arXiv:1612.04460*. [Online]. Available: https://arxiv.org/abs/1612.04460

[20] H.-S. Chang, Z. Wang, L. Vilnis, and A. McCallum, "Distributional inclusion vector embedding for unsupervised hypernymy detection," Oct. 2017, *arXiv:1710.00880*. [Online]. Available: https://arxiv.org/abs/1710.00880

[21] C. Wang, X. He, and A. Zhou, "A short survey on taxonomy learning from text corpora: Issues, resources and recent advances," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark, Sep. 2017, pp. 1190–1203.

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," Jan. 2013, *arXiv:1301.3781*. [Online]. Available: https://arxiv.org/abs/1301.3781

[23] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Doha, Qatar, Oct. 2014, pp. 1532–1543.

[24] M. Baroni, R. Bernardi, N. Q. Do, and C. C. Shan, "Entailment above the word level in distributional semantics," in *Proc. Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2012, pp. 23–32.

[25] R. Fu, J. Guo, B. Qin, W. Che, H. Wang, and T. Liu, "Learning semantic hierarchies via word embeddings," in *Proc. Annu. Meeting Assoc. for Comput. Linguistics*, Baltimore, MD, USA, Jun. 2014, pp. 1199–1209.

[26] P. D. Turney and S. M. Mohammad, "Experiments with three approaches to recognizing lexical entailment," *Natural Lang. Eng.*, vol. 21, no. 3, pp. 437–476, May 2015.

[27] K. A. Nguyen, M. Köper, S. S. I. Walde, and N. T. Vu, "Hierarchical embeddings for hypernymy detection and directionality," Jul. 2017, *arXiv:1707.07273*. [Online]. Available: https://arxiv.org/abs/1707.07273

[28] Z. Yu, H. Wang, X. Lin, and M. Wang, "Learning term embeddings for hypernymy identification," in *Proc. Int. Joint Conf. Artif. Intell.*, Buenos Aires, Argentina, Jul. 2015, pp. 1290–1397.

[29] S. Roller and K. Erk, "Relations such as hypernymy: Identifying and exploiting hearst patterns in distributional vectors for lexical entailment," May 2016, *arXiv:1605.05433*. [Online]. Available: https://arxiv.org/abs/1605.05433

[30] L. Tan, R. Gupta, and J. van Genabith, "USAAR-WLV: Hypernym generation with deep neural nets," in *Proc. Int. Workshop Semantic Eval.*, Denver, CO, USA, Jun. 2015, pp. 932–937.

[31] J. Yamane, T. Takatani, H. Yamada, M. Miwa, and Y. Sasaki, "Distributional hypernym generation by jointly learning clusters and projections," in *Proc. Int. Conf. Comput. Linguistics, Tech. Papers*, Osaka, Japan, Dec. 2016, pp. 1871–1879.

[32] C. Wang and X. He, "Chinese hypernym-hyponym extraction from user generated categories," in *Proc. Int. Conf. Comput. Linguistics, Tech. Papers*, Osaka, Japan, Dec. 2016, pp. 1350–1361.

[33] C. Wang, Y. Fan, X. He, and A. Zhou, "A family of fuzzy orthogonal projection models for monolingual and cross-lingual hypernymy prediction," in *Proc. World Wide Web Conf.*, May 2019, pp. 1965–1976.

[34] R. Fu, B. Qin, and T. Liu, "Exploiting multiple sources for open-domain hypernym discovery," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Seattle, WA, USA, Oct. 2013, pp. 1224–1234.

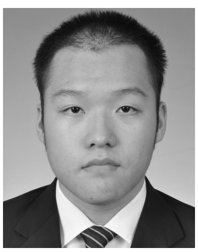
[35] J. Li, C. Wang, X. He, R. Zhang, and M. Gao, "User generated content oriented chinese taxonomy construction," in *Proc. Asia–Pacific Web Conf.*, Guangzhou, China, Sep. 2015, pp. 623–634.

[36] O. Levy, S. Remus, C. Biemann, and I. Dagan, "Do supervised distributional methods really learn lexical inference relations?" in *Proc. Conf. North Amer. Chapter Assoc. for Comput. Linguistics, Hum. Lang. Technol.*, Denver, CO, USA, May 2015, pp. 970–976.

[37] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," Sep. 2014, *arXiv:1406.1078*. [Online]. Available: https://arxiv.org/abs/1406.1078

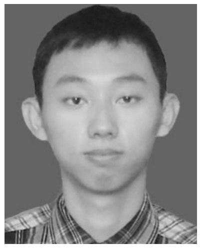
[38] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 3856–3866.

[39] K. Donnelly, "SNOMED-CT: The advanced terminology and coding system for ehealth," *Stud. Health Technol. Informat.*, vol. 121, no. 121, p. 279, 2006.
[40] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," Dec. 2012, *arXiv:1212.5701*. [Online]. Available: https://arxiv.org/abs/1212.5701
[41] Y. Liu, Y. Zhou, S. Wen, and C. Tang, "A strategy on selecting performance metrics for classifier evaluation," *Int. J. Mobile Comput. Multimedia Commun.*, vol. 6, no. 4, pp. 20–35, Apr. 2014.
[42] Y. Zhou and Y. Liu, "Correlation analysis of performance metrics for classifier," in *Proc. Int. FLINS Conf. Decision Making Soft Comput.*, João Pessoa, Brazil, Sep. 2014, pp. 487–492.
[43] T. Jiang, M. Liu, B. Qin, and T. Liu, "Constructing semantic hierarchies via fusion learning architecture," in *Proc. China Conf. Inf. Retr.*, Shanghai, China, Jul. 2017, pp. 136–148.
[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.
[45] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," Sep. 2014, *arXiv:1409.0473*. [Online]. Available: https://arxiv.org/abs/1409.0473
[46] M. Baroni and A. Lenci, "How we blessed distributional semantic evaluation," in *Proc. GEMS Workshop GEometrical Models Natural Lang. Semantics*, Edinburgh, Scotland, Jul. 2011, pp. 1–10.
[47] E. Santus, A. Lenci, T.-S. Chiu, Q. Lu, and C.-R. Huang, "Nine features in a random forest to learn taxonomical semantic relations," Mar. 2016, *arXiv:1603.08702*. [Online]. Available: https://arxiv.org/abs/1603.08702
[48] E. Santus, F. Yung, A. Lenci, and C. Huang, "Evalution 1.0: An evolving semantic dataset for training and evaluation of distributional semantic models," in *Proc. Workshop Linked Data Linguistics, Resour. Appl.*, Beijing, China, Jul. 2015, pp. 64–69.



CHENMING XU received the bachelor's degree from the East China University of Science and Technology, Shanghai, China, in 2018, where he is currently pursuing the master's degree. His research interests include information extraction and text mining.

YANGMING ZHOU received the Ph.D. degree from the University of Angers, Angers, France, in 2018. He is currently a Lecturer with the School of Information Science and Engineering, East China University of Science and Technology, China. His current research interests include natural language processing, machine learning, and evolutionary computation.



QI WANG received the bachelor's degree from the East China University of Science and Technology, Shanghai, China, in 2016, where he is currently pursuing the master's degree. His research interests include information extraction and text mining.

ZHIYUAN MA received the Ph.D. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2019. He is currently a Postdoctoral Fellow with the School of Information Science and Engineering, East China University of Science and Technology, China. His current research interests include information extraction and text mining.

YAN ZHU received the Ph.D. degree from Paris-Sud University, Paris, France, in 2010. She is currently an Associate Professor with the School of Science, East China University of Science and Technology, China. Her current research interest includes graph theory and it's applications.