Complex & Intelligent Systems (2023) 9:6065–6085
https://doi.org/10.1007/s40747-023-01064-w

ORIGINAL ARTICLE

A bilateral context and filtering strategy-based approach to Chinese entity synonym set expansion

Subin Huang[1] · Yu Xiu[1] · Jun Li[1] · Sanmin Liu[1] · Chao Kong[1]

Received: 11 October 2022 / Accepted: 22 March 2023 / Published online: 25 April 2023

Abstract

Entity synonyms play a significant role in entity-based tasks. Previous approaches use linguistic syntax, distributional, and semantic features to expand entity synonym sets from text corpora. Due to the flexibility and complexity of the Chinese language expression, the aforementioned approaches are still difficult to expand entity synonym sets robustly from Chinese text, because these approaches fail to track holistic semantics among entities and suffer from error propagation. This paper introduces an approach for expanding Chinese entity synonym sets based on bilateral context and filtering strategy. Specifically, the approach consists of two novel components. First, a bilateral-context-based Siamese network classifier is proposed to determine whether a new entity should be inserted into the existing entity synonym set. The classifier tracks the holistic semantics of bilateral contexts and is capable of imposing soft holistic semantic constraints to improve synonym prediction. Second, a filtering-strategy-based set expansion algorithm is presented to generate Chinese entity synonym sets. The filtering strategy enhances semantic and domain consistencies to filter out wrong synonym entities, thereby mitigating error propagation. Experimental results on two Chinese real-world datasets demonstrate that the proposed approach is effective and outperforms the selected existing state-of-the-art approaches to the Chinese entity synonym set expansion task.

Keywords Synonym set expansion · Siamese network · Bilateral context · Filtering strategy

Introduction

Mining entity synonym set is an important task for many entity-based downstream applications, such as knowledge graph construction [1–4], taxonomy learning [5–8], and question answering [9–11]. An entity synonym set usually contains several different strings representing an identical entity [12–14]. For example, English strings {“The United States”, “America”, “USA”} are the alternative ways to represent the real entity “The United States of America” and Chinese strings {“洋芋”, “土豆”} are the alternative ways to represent the real entity “马铃薯 (potato)”. Take the question “Do you need a visa for the USA?” as an example, understanding “USA” refers to a country “The United States of America” is crucial for an artificial intelligence system to satisfy the user information need [15].

The existing approaches use linguistic syntax, distributional, and semantic features to expand entity synonym sets from English text corpora. These approaches can be grouped into four categories: pattern-based approaches [16–19], distribution-based approaches [20–25], graph-based approaches [26–29], and two-step approaches [15, 30–32].

However, due to the flexibility and complexity of the Chinese language expression, it is still difficult to expand entity synonym sets robustly from Chinese text [33]. From a linguistic point of view, Chinese is an ideogram language with complex and irregular grammar, lexical structure, and semantics. For example, Chinese has no specific tenses and voices and no distinction between singular and plural forms. In addition, the word order of Chinese is significantly different from that of English, and there are no spaces between words in Chi-

✉ Sanmin Liu
sanmin.liu@ahpu.edu.cn

Subin Huang
subinhuang@ahpu.edu.cn

Yu Xiu
xiuyu@ahpu.edu.cn

Jun Li
edmondlee@ahpu.edu.cn

Chao Kong
kongchao@ahpu.edu.cn

[1] School of Computer and Information, Anhui Polytechnic University, Wuhu 241000, People’s Republic of China

nese [6]. Therefore, the aforementioned approaches are not necessarily suitable for expanding entity synonym sets from Chinese text and have the following limitations:

- Weak holistic semantic inference: Distribution-based approaches usually consider distributional statistics and similarity information, and graph-based approaches often use clustering algorithms. This may lead to the loss of holistic semantics among entities.
- Unsatisfactory robustness: Manual/semimanual and pattern-based approaches achieve relatively high precision but low recall, whereas distribution-based approaches achieve relatively high recall but low precision.
- Error propagation: Graph-based and two-step approaches often suffer from error propagation. This is because the available resources used by graph-based approaches and the first task of the two-step approaches are not all correct. These issues lead to the propagation of errors in subsequent processing.
- Lack of labeled training datasets: Distribution-based and two-step approaches usually require labeled training datasets to train a detection model. However, the labeled Chinese entity synonym set training datasets are not always available and are expensive to develop.

In this work, we propose a bilateral context and filtering strategy approach to mitigate the aforementioned limitations and improve the expansion of Chinese entity synonym sets. Specifically, the approach first obtains a large-scale Chinese entity vocabulary using a Chinese knowledge base and applies an entity linker to acquire distant supervision knowledge. Second, a bilateral-context-based Siamese network classifier is developed to evaluate an input Chinese entity for its inclusion into the existing synonym set. The classifier tracks the holistic semantics of bilateral contexts and is capable of imposing soft holistic semantic constraints to improve synonym prediction. Third, an entity synonym set expansion algorithm combined with the bilateral-context-based Siamese network classifier and an entity expansion filtering strategy is used to expand the Chinese entity synonym sets. The filtering strategy consists of similarity filtering and domain filtering. The strategy is capable of enhancing semantic and domain consistencies to filter out wrong Chinese synonym entities and mitigate the problem of error propagation caused by the Siamese network classifier.

The main contributions of this study are threefold:

- We propose a bilateral-context-based Siamese network classifier to track Chinese synonyms.
- We propose a filtering-strategy-based set expansion algorithm to expand Chinese entity synonym sets.
- Two Chinese real-world entity synonym set expansion datasets are constructed. The datasets and the source code of our approach are available at https://github.com/huangsubin/CNSynSetE.

The proposed approach is applied to two Chinese real-world entity synonym set datasets. A detailed experimental analysis and evaluation of the proposed approach is performed and the results are compared with those of selected state-of-the-art existing approaches. The results demonstrate that the proposed approach is effective and outperforms the existing state-of-the-art approaches used for the Chinese entity synonym set expansion task. In addition, the ablation and case studies demonstrate that the bilateral context and the entity filtering strategy play a significant role in improving the performance of the proposed approach.

The remainder of this paper is structured as follows. A brief overview of the existing synonym set discovery approaches and the main features of the proposed approach are presented in the section “Related works”. The details of the proposed approach are discussed in the section “Materials and methods”. Experimental results are presented and analyzed in the section “Experiments”. Conclusions and directions for future studies are presented in the section “Conclusion”.

Related works

Various approaches to discover synonym sets from text using intelligent technologies (e.g., data mining and deep learning) are available. These approaches can be grouped into four types: pattern-based, distribution-based, graph-based, and two-step approaches.

Pattern-based approaches

The pattern-based approach was first proposed by Hearst [34]. Such an approach uses predefined lexical patterns (e.g., N_1 such as N_2, where N_1 and N_2 denote nouns or noun phrases) to acquire hyponyms [6]. Following [34], some researchers employed pattern-based approaches to discover synonym sets from text. For example, based on the predefined lexical pattern (e.g., N_3 refers to N_4, where N_3 and N_4 denote nouns or noun phrases), the pattern-based approach can infer that $\{N_3, N_4\}$ is a synonym set.

However, manually predefining the synonymous lexical patterns is time-consuming and laborious. Therefore, subsequent studies were devoted to automatically acquiring synonymous lexical patterns from corpora. McCrae and Collier [16] first presented an approach to discover synonym patterns. Next, they used the generated patterns to build synonym feature vectors and exploited logistic regression to predict synonym sets. Wang et al. [17] presented a pattern-construction method to mine verb synonyms and antonyms. They used multiple patterns to improve the recall of verb synonyms and antonyms. Nguyen et al. [19] proposed a pattern-based neural network approach to discover synonyms. They used lexical patterns extracted from the syntactic parse trees and exploited the distance of the syntactic path to capture new patterns.

For the Chinese language, Kwong and Tsou [35] used lexical items to extend and enhance Tongyici Cilin (a Chinese synonym dictionary). Li and Lu [18] proposed a hybrid mining approach to discover Chinese noun/verb synonym sets. They used syntactic patterns and semantic knowledge to improve the performance of Chinese noun/verb synonym set extraction.

Distribution-based approaches

Based on the distributional hypothesis [20], distribution-based approaches use distributional statistical features to mine synonym sets. These approaches consider that words presented in identical or similar contexts are more likely to be synonyms. For instance, the synonymous words "United State" and "USA" often appear in identical or similar contexts. Distribution-based approaches represent words using distributional statistical features and use these features to learn whether the given words are synonymous [21]. These approaches usually discover synonym sets based on the existing synonym seeds.

Turney [22] proposed an unsupervised approach for discovering synonyms. They used information retrieval (IR) and pointwise mutual information (PMI) to decide whether the given words are synonymous. Chakrabarti et al. [23] presented a general framework for robustly mining synonyms. They used the pseudo-document similarity function and query context similarity to capture the synonymous features and used the MapReduce technology to discover synonyms from large-scale Web text. Qu et al. [21] proposed an automatic method for discovering synonyms from domain-specific text. They used corpus-level distributional features and textual patterns to enhance the synonym signals of distant supervision. Zhang et al. [25] proposed a synonym discovery approach using a distributional-hypothesis-based multicontext setting. They presented a neural network model with multiple pieces of contexts to learn whether two given entities are synonymous.

For the Chinese language, Yu et al. [36] evaluated the performance of two distributional statistical features, namely PMI and a 5-gram language models, on Chinese synonym choice and reported that the 5-gram language model outperformed the PMI in terms of accuracy. Gan [37] studied the collocations of Chinese synonym sets. They used the distributional statistic feature named mutual information (MI) to analyze the collected corpus and the Chinese synonyms from the aspects of prosody, register, and semantic features. Ma et al. [24] proposed a Chinese synonym extraction approach based on multidistribution features. They used three distribution features to score the candidate synonym sets and regarded synonym extraction as a ranking task.

Graph-based approaches

Graph-based approaches first build a graph in which nodes denote entities and edges denote the relationship between the entities. Next, these approaches use a clustering algorithm to induce synonyms from the graph [28]. Generally, the graph is built from available resources (e.g., Web link text or Wiktionary).

Dorow and Widdows [26] proposed an unsupervised method for learning word sense. They used a graph model to represent words and their relationships. Furthermore, they used a Markov clustering algorithm to discover synonym sets from the graph. Based on word embeddings and synonym dictionaries, Ustalov et al. [28] proposed a weighted graph-based synonym discovery approach. First, they constructed a weighted synonym graph from available resources and used word sense induction to process ambiguous words. Next, they used a meta-clustering algorithm to discover synonym sets from the weighted graph. Ercan and Haziyev [29] presented an automatic synonym construction approach based on a translation graph. They built a translation graph using multiple Wiktionaries and used clustering, greedy, and supervised learning algorithms to discover synonyms from the translation graph.

For the Chinese language, Lu and Hou [38] proposed a Chinese synonym-acquiring approach using a wiki repository. They constructed an associated word graph using relational links extracted from the wiki repository and used the PageRank algorithm to mine synonyms from the associated word graph. Duan et al. [27] proposed a sememe-tree-based approach for reducing Chinese synonyms. They used the distances between words in the sememe tree to reduce the synonyms.

Two-step approaches

Two-step approaches deploy two sequential subtasks to discover the entity synonym sets. Such approaches first train a synonym prediction model to determine whether the given candidate string pairs are synonymous. Subsequently, the approach uses a synonym expansion algorithm combined with the above prediction model to acquire the synonym sets. These approaches are usually capable of extracting semantic relations among candidate strings and grouping all synonyms together from candidate strings [13, 15].

Ren and Cheng [30] proposed an approach including a heterogeneous graph-based data model and a graph-based ranking algorithm to discover synonyms from web text. They exploited string names, some important structured attributes, subqueries, and tailed web pages to acquire more synonyms. Shen et al. [31] proposed an approach including a context feature selection method and a ranking-based ensemble model to mine synonym sets from free-text corpora. Shen et al. [15] presented an efficient entity synonym set generation approach to mine entity synonym sets. They constructed a set-instance classifier to determine whether given candidate string pairs are synonymous and used a set generation algorithm to expand entity synonym sets.

For the Chinese language, Huang et al. [32] proposed an approach including extraction and cleaning steps for generating Chinese entity synonym sets. In the extraction step, they used direct extraction, pattern-based extraction, and neural mining extraction to obtain candidate Chinese entity synonym sets. In the cleaning step, they used lexical and semantic rules, domain filtering, and similarity filtering to improve the accuracy of the obtained Chinese entity synonym sets.

Discussion

In the above subsection, four synonym set discovery approaches are reviewed. Here, the main features of our approach and the approaches reviewed above will be discussed in detail.

Pattern-based approaches can achieve relatively high accuracy. However, these approaches often suffer from low coverage. In contrast, distribution-based approaches usually achieve relatively high coverage but have low accuracy. Graph-based approaches usually use a clustering algorithm (e.g., Markov clustering and PageRank algorithms) to discover synonyms. However, these approaches lose synonymous semantics in graph-based synonym clustering, which renders the accuracy and coverage of synonym mining unsatisfactory. Most two-step approaches are supervised. Such approaches require labeled synonym datasets. However, labeled synonym datasets are not always available and are expensive to develop.

The bilateral context and filtering strategy-based approach proposed herein achieves a large-scale Chinese entity vocabulary based on the Chinese knowledge base and applies an entity linker to generate Chinese entity synonym set datasets using the Chinese encyclopedia. To capture more synonymous semantics, a bilateral-context-based Siamese network classifier is proposed to determine whether a new input Chinese entity should be inserted into an existing synonym set. The holistic association semantics of bilateral contexts among entities are fused into this classifier, which is capable of imposing soft holistic semantic constraints for determining whether a new input Chinese entity instance should be inserted into an existing synonym set to improve the robustness of the classifier. In the entity synonym set expansion algorithm, the proposed approach presents an entity expansion filtering strategy for filtering incorrect Chinese synonym entities, thereby mitigating the problem of error propagation.

Materials and methods

This section introduces the definitions of the concepts involved and the framework of the proposed approach. Furthermore, it includes a detailed discussion on its framework and components.

Definitions and problem statement

First, we introduce some important concepts.

- **Synonym**. Synonyms are strings or words that have the same or almost the same meaning in a language [15]. Synonyms are ubiquitous in all human natural languages. For example, “USA” and “United States” refer to the same country; “Abuse” and “Maltreatment” mean cruel or inhumane treatment.
- **Entity synonym set**. An entity synonym set denotes a group of strings or words that represents an identical or similar entity in a language. For example, {“The United Kingdom”, “Britain”, “U.K.”} is an entity synonym set, because the strings in the set denote the same country: “United Kingdom of Great Britain and Northern Ireland”.
- **Knowledge base**. A knowledge base contains many entities and facts [15, 21]. This study focuses on exploiting the entities in the Chinese knowledge base to acquire Chinese entity synonym set datasets from the Chinese encyclopedia.
- **Problem statement**. Given a Chinese text corpus C and a vocabulary V generated from C, the objective of this study is to expand Chinese entity synonym sets from V based on the clues (e.g., bilateral context and filtering features) mined from C and V. Actually, the entity synonyms are transitive (entities cannot be multivocal words) and symmetric.
  Transitive: $(a \overset{\text{syn}}{\rightarrow} b \wedge b \overset{\text{syn}}{\rightarrow} c) \Rightarrow (a \overset{\text{syn}}{\rightarrow} c)$.
  Symmetric: $(a \overset{\text{syn}}{\rightarrow} b) \Rightarrow (b \overset{\text{syn}}{\rightarrow} a)$.

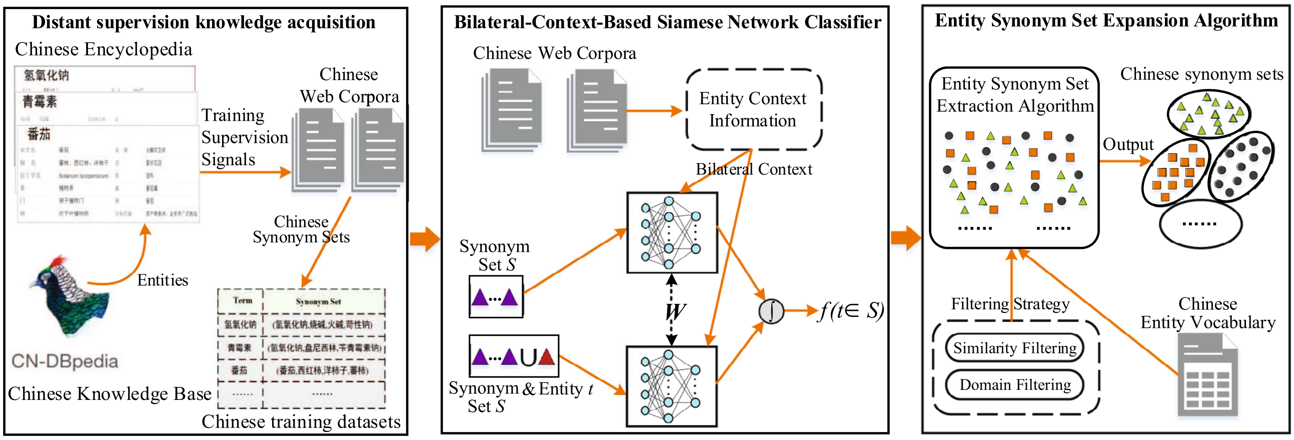

Distant supervision knowledge acquisition
Chinese Encyclopedia
氢氧化钠
青霉素
番茄
Training Supervision Signals
Chinese Web Corpora
Chinese Synonym Sets
Entities
CN-DBpedia
Chinese Knowledge Base
Term
Synonym Set
氢氧化钠
(氢氧化钠,烧碱,火碱,苛性钠)
青霉素
(氢氧化钠,盘尼西林,苄青霉素钠)
番茄
(番茄,西红柿,洋柿子,蕃柿)
......
......
Chinese training datasets
Bilateral-Context-Based Siamese Network Classifier
Chinese Web Corpora
Entity Context Information
Bilateral Context
Synonym Set S
W
f(t∈ S)
Synonym & Entity t Set S
Entity Synonym Set Expansion Algorithm
Entity Synonym Set Extraction Algorithm
Chinese synonym sets
Output
Filtering Strategy
Similarity Filtering
Domain Filtering
Chinese Entity Vocabulary

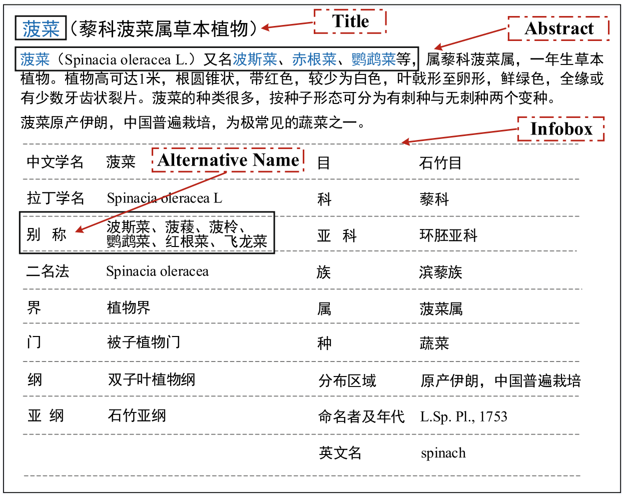
Fig. 1 Framework of the proposed approach to Chinese entity synonym set expansion

Here, a, b and c are the strings or words in V. Relation $a \overset{syn}{\to} b$ denotes that a and b are synonymous. Therefore, in an entity synonym set, all entities are synonymous with each other.

Overview of framework

As depicted in Fig. 1, the framework of the proposed approach to Chinese entity synonym set expansion comprises three components: distant supervision knowledge acquisition, bilateral-context-based Siamese network classifier, and entity synonym set expansion algorithm.

- Distant supervision knowledge acquisition: This component entails obtaining the Chinese entity vocabulary from the Chinese knowledge base and acquiring entity synonym set datasets from Chinese web corpora using the Chinese encyclopedia as training supervision signals.
- Bilateral-context-based Siamese network classifier: A classifier is built to determine whether a new input Chinese entity should be inserted into the existing Chinese entity synonym set. The classifier contains a Siamese network with entity bilateral context and is capable of learning more synonymous features.
- Entity synonym set expansion algorithm: A filtering-strategy-based set expansion algorithm is designed to expand Chinese entity synonym sets. The algorithm is combined with the bilateral-context-based Siamese network classifier and entity expansion filtering strategy to improve the performance of the Chinese entity synonym set expansion task.



菠菜（藜科菠菜属草本植物）

菠菜（Spinacia oleracea L.）又名波斯菜、赤根菜、鹦鹉菜等，属藜科菠菜属，一年生草本植物。植物高可达1米，根圆锥状，带红色，较少为白色，叶戟形至卵形，鲜绿色，全缘或有少数牙齿状裂片。菠菜的种类很多，按种子形态可分为有刺种与无刺种两个变种。

菠菜原产伊朗，中国普遍栽培，为极常见的蔬菜之一。

中文学名	菠菜	目	石竹目
拉丁学名	Spinacia oleracea L	科	藜科
别　称	波斯菜、菠薐、菠柃、鹦鹉菜、红根菜、飞龙菜	亚　科	环胚亚科
二名法	Spinacia oleracea	族	滨藜族
界	植物界	属	菠菜属
门	被子植物门	种	蔬菜
纲	双子叶植物纲	分布区域	原产伊朗，中国普遍栽培
亚　纲	石竹亚纲	命名者及年代	L.Sp. Pl., 1753
		英文名	spinach

Fig. 2 Entity synonyms in Chinese encyclopedia

Distant supervision knowledge acquisition

A knowledge base consists of many entities and facts. Such entities and facts are capable of providing distant supervision signals to mine more entity synonym sets from raw text corpora [21, 39]. We use a Chinese knowledge base, named CN-Dbpedia [2], to construct Chinese entity vocabulary V and then automatically obtain a collection of Chinese synonym sets from the Chinese encyclopedia named Baidu Encyclopedia. As depicted in Fig. 2, some Chinese entity synonyms are available from Baidu Encyclopedia. For example, if a Chinese entity mention has synonyms, there is an infobox named “别称 (alternative name)” for enumerating the synonym entities of the entity mention [32]. Therefore, we merge the entity mention into the enumerated entity synonyms and obtain a collection of Chinese entity synonym sets named CN-SynSets. The following rule is employed to reduce the

Algorithm 1 Set-entity pair generation algorithm

```
Input: (1) A synonym set dataset $SynSet = \{BDSynSetTra\ or\ SGSynSetTra\}$; (2) a Negative sample number $k$.
Output: $SEPs$: a collection of set-entity pairs for $SynSet$
Initial set-entity pair collection, let $SEPs \rightarrow \{\ \}$;
for each $Syn_i \in SynSet$ do
    for each $t_i \in Syn_i$ do
        Generate a positive set-term pair $pos_j = \{(t_1, \cdots, t_{i-1}, \cdots, t_{i+1}, \cdots, t_n); t_i; 1\}$
        Randomly sample $k$ entities $t_i^{neg}|_1^k$ from $SynSet$ but not mentioned in $Syn_i$
        Generate $k$ negative set-term pairs for $pos_j$, denoted as $neg_j|_1^k = \{(t_1, \cdots, t_{i-1}, \cdots, t_{i+1}, \cdots, t_n); t_i^{neg}|_1^k; 0\}$
        Merge $pos_j$ and $neg_j|_1^k$ into $SEP \rightarrow \{(pos_j, neg_j|_1^k)\}$;
    end for
end for
return $SEPs$
```

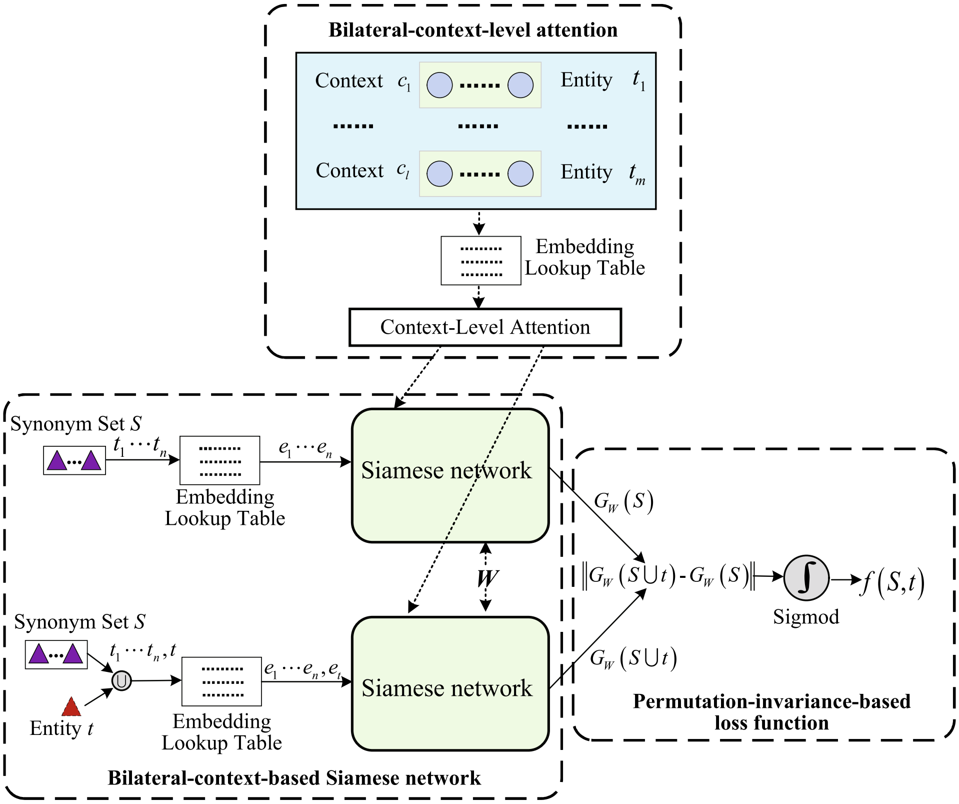


Fig. 3 Architecture of bilateral context-based Siamese network classifier

merging errors: each term in the enumerated entity synonyms must be a named entity, noun, or noun phrase.

To generate Chinese distant supervision knowledge, we apply [1] Hanlp[2] to link the entities involved in Cn-SynSets to two Chinese web corpora, namely, Baidu Encyclopedia articles[1] and SogouCA.[3] Next, we generate two real-world Chinese entity synonym set datasets: BDSynSetTra and SGSynSetTra, respectively, denoted as

$$\text{BDSynSetTra} = \{\text{Syn}_1, \ldots, \text{Syn}_i, \ldots, \text{Syn}_N\}, \quad (1)$$

$$\text{SGSynSetTra} = \{\text{Syn}_1, \ldots, \text{Syn}_i, \ldots, \text{Syn}_M\}, \quad (2)$$

where $\text{Syn}_i = \{t_1, \ldots, t_i, \ldots, t_n\}$ denotes a synonym set in BDSynSetTra and SGSynSetTra, N and M denote numbers of the synonym sets, n denotes the number of entities in a synonym set, and t_i denotes the entities involved in synonym sets.

We build a collection of set-entity pairs (SEPs) for BDSynSetTra and SGSynSetTra using Algorithm 1. The impact of negative sample size k on the bilateral-context-based Siamese network classifier is discussed in the section "Effect of negative sample size".

[1] https://baike.baidu.com/.

[2] https://github.com/hankcs/HanLP.

[3] http://www.sogou.com/labs/resource/ca.php.

Bilateral-context-based Siamese network classifier

Shen et al. [15] proposed a synonym set-instance classifier to determine whether a new input entity should be inserted into the existing synonym set. They designed set scorer $q(S)$ to obtain a score for an input entity synonym set S. Next, they used set scorer $q(\bar{S})$ to obtain a new score for a new input entity synonym set $\bar{S}$, where $\bar{S} = S \cup \{t\}$ and t is a new entity. The objective function of the set-instance classifier proposed in [15] is to use a sigmoid function to convert the difference between $q(\bar{S})$ and $q(S)$ to a probability

$$p(t \in S) = \text{Sigmoid}(q(\bar{S}) - q(S)). \tag{3}$$

However, Shen et al. [15] use only entity embedding to capture synonym signals, ignoring the context semantics between the new entities and original synonym sets. These context semantics are capable of imposing soft constraints for determining whether a new input Chinese entity instance should be inserted into an existing synonym set to improve the performance of Chinese entity synonym set expansion [40].

We present a bilateral-context-based Siamese network classifier to capture more semantics to improve the performance of Chinese entity synonym set-instance prediction. The difference between the bilateral-context-based Siamese network and the Siamese network is that we use not only entity embeddings but also the holistic semantic association of bilateral contexts between entities. Therefore, the bilateral-context-based Siamese network classifier is able to track more holistic synonym signals to determine whether a new input Chinese entity should be included in the existing synonym set.

As depicted in Fig. 3, the architecture of the bilateral-context-based Siamese network classifier bc-snc(S, t) comprises three components: bilateral-context-level attention, bilateral-context-based Siamese network, and permutation-invariance-based loss function. Given synonym set $S = \{t_1, \ldots, t_n\}$ and a new input entity t, classifier bc-snc(S, t) learns the hidden representations $G_w(S)$ and $G_w(S \cup t)$ using the bilateral-context-based Siamese network. Subsequently, bc-snc(S, t) calculates the difference between $G_w(S)$ and $G_w(S \cup t)$ and uses a permutation-invariance-based loss function to determine whether new entity t should be included in synonym set S.

Bilateral-context-level attention

(i) Generating bilateral context

The properties of the contexts that surround entities are highly useful for expanding entity synonyms [25]. For example, “乙酰水杨酸” (“acetylsalicylic acid”) and “阿司匹林” (“aspirin”) are synonymous. The contexts such as “消炎” (“anti-inflammatory”) and “镇痛” (“analgesia”) surrounding these two entities are similar. Based on the distributional hypothesis [21], we observe that the synonymous entities are related to the following factor.

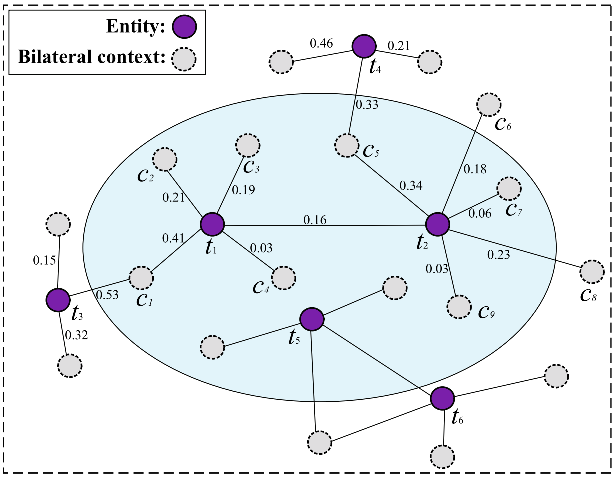


Fig. 4 Context association semantic network

Observation 1 *Context semantic consistency: If two entities are synonymous, they are more likely to be mentioned in the same context.*

Observation 1 guides us to consider the context of entities to capture more synonymous features. We used a sliding window method to generate the bilateral context of the entities.

Given entity t_i, a sentence u_j that contains entity t_i, and a window size of $d = 5$, the bilateral context of t_i is acquired from the five words $\{t_{i-5}, t_{i-4}, t_{i-3}, t_{i-2}, t_{i-1}\}$ that precede it and five words $\{t_{i+1}, t_{i+2}, t_{i+3}, t_{i+4}, t_{i+4}\}$ that follow it.

After acquiring the bilateral context of entities, a context association semantic network is constructed. The objective of constructing a context association semantic network is to establish the relationship between entities and bilateral context. As depicted in Fig. 4, the context association semantic network contains entities, bilateral contexts, and weighted paths between entities and bilateral contexts. Weighted path wp(t_i, c_j) is defined as follows:

$$\text{wp}(t_i, c_j) = \frac{f(t_i, c_j)}{\sum_{j=1}^{m} f(t_i, c_j)}, \tag{4}$$

where t_i is an entity and c_j denotes its bilateral context, $f(t_i, c_j)$ is the number of cooccurrences between t_i and c_j, and m denotes the number of bilateral contexts for entity t_i. Given entity t_i, we retrieve its n-hop bilateral contexts using the top-k weighted paths. For example, setting $n = 2$ and $k = 4$, based on Fig. 4, the bilateral contexts of t_1 and t_2 are $\{c_1, c_2, c_3, t_3\}$ and $\{c_5, t_4, c_8, c_6\}$, respectively.

(ii) Building context-level attention

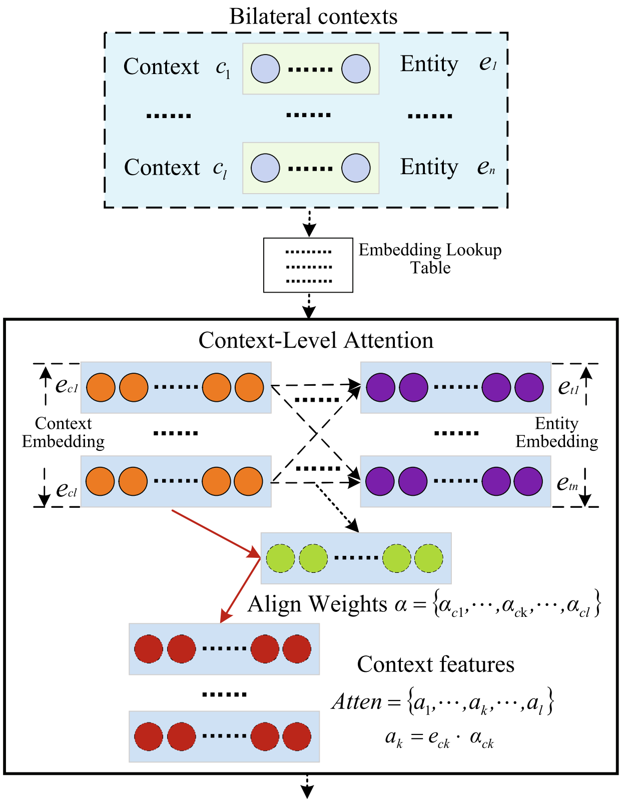


Fig. 5 Context-level attention mechanism

A context-level attention mechanism is proposed to learn lower weights for weakly relevant bilateral contexts and higher weights for strongly relevant contexts. Mikolov et al. [41] studied the relationships between some word embeddings, such as $e_{\text{London}} - e_{\text{Beijing}} \approx e_{\text{Paris}} - e_{\text{Tokyo}}$ and $e_{\text{USA}} - e_{\text{United States}} \approx e_{\text{U.K}} - e_{\text{United Kingdom}}$, which shows that the relationships between different words can be reflected in their word embeddings. Based on the above ideas, Ji et al. [42] used $e_{\text{relation}} = e_i - e_j$ to represent the relationship between word embeddings e_i and e_j and used the similarity between e_{relation} and a given instance expression to learn the attention weight.

Figure 5 depicts the detailed structure of the proposed attention mechanism. Given input entity set $S = \{t_1, \ldots, t_n\}$ and its bilateral context set $C = \{c_1, \ldots, c_l\}$, the details of context-level attention are as follows:

- First, context-level attention transforms $S = \{t_1, \ldots, t_n\}$ and $C = \{c_1, \ldots, c_l\}$ into embedding sets $e_s = \{e_{t1}, \ldots, e_{tn}\}$ and $e_c = \{e_{c1}, \ldots, e_{cl}\}$, respectively.
- Second, for each context embedding $e_{ck} \in e_c$, context-level attention calculates the align weights $\alpha = \{\alpha_{c1}, \ldots, \alpha_{ck}, \ldots, \alpha_{cl}\}$, denoted by

$$\alpha_{ck} = \frac{\sum_{j=1}^{n} e_{ck} \cdot e_{tj}}{\sum_{i=1}^{l} \sum_{j=1}^{n} e_{ci} \cdot e_{tj}}. \tag{5}$$

- Third, the outputs of context-level attention are context features Atten $= \{a_1, \ldots, a_k, \ldots, a_l\}$, where $a_k = e_{ck} \cdot \alpha_{ck}$.

Bilateral-context-based Siamese network

The bilateral-context-based Siamese network is designed to capture synonymous features from the bilateral context of entities and entity embeddings. As depicted in Fig. 6, this network includes two components: the embedding feature extractor and bilateral-context-based feature extractor.

(i) Embedding feature extractor

The objective of the embedding feature extractor is to extract synonymous features from entity embeddings. As depicted in Fig. 6, the architecture of the embedding feature extractor is as follows:

- First, given input set $S = (t_1, \ldots, t_m)$, $S \in \{(t_1, \ldots, t_n), (t_1, \ldots, t_n, t)\}$, where m is equal to n or $n + 1$. The embedding feature extractor represents set $(t_1, \ldots, t_m)$ as embeddings $(e_1, \ldots, e_m)$ via the embedding lookup table.
- Second, embeddings $(e_1, \ldots, e_m)$ are input into neural network $\theta_1(\cdot)$ with a two-layer fully connected structure. Next, embeddings $(e_1, \ldots, e_m)$ are transformed into m hidden representations as $H_1 = (\theta_1(e_1), \ldots, \theta_1(e_m))$.
- Third, a summation operation is used to change H_1 into a hidden representation $H_2 = \sum_{i=1}^{m} \theta_1(e_i)$.
- Fourth, H_2 is input into neural network $\theta_2(\cdot)$ with a three-layer fully connected structure. Subsequently, representation H_2 is transformed into a hidden representation as $H_3 = \theta_2(H_2)$.

(ii) Bilateral-context-based feature extractor

The objective of the bilateral-context-based feature extractor is to capture synonymous features from context features Atten $= \{a_1, \ldots, a_k, \ldots, a_l\}$ generated from the context-level attention mechanism. As depicted in Fig. 6, the structure of the bilateral-context-based feature extractor is as follows.

- First, context features Atten $= \{a_1, \ldots, a_k, \ldots, a_l\}$ are input into neural network $\bar{\theta}_1(\cdot)$ with a two-layer fully connected structure. Next, Atten $= \{a_1, \ldots, a_k, \ldots, a_l\}$ are transformed into l hidden representations as $\bar{H}_1 = (\bar{\theta}_1(a_1), \ldots, \bar{\theta}_1(a_l))$.
- Second, a summation operation is used to change $\bar{H}_1$ into hidden representation $\bar{H}_2 = \sum_{i=1}^{l} \bar{\theta}_1(a_i)$.

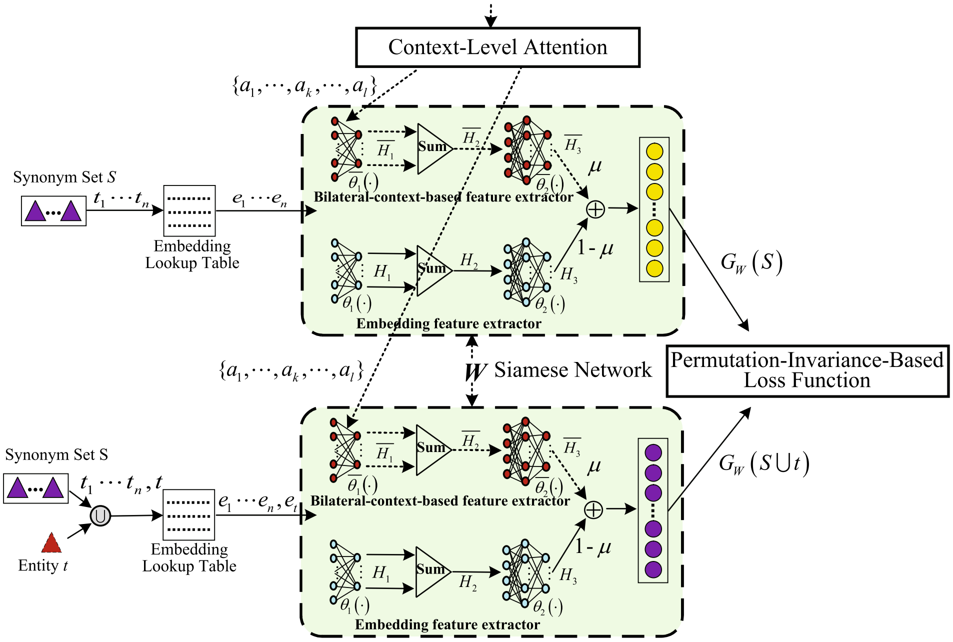


Fig. 6 Architecture of bilateral-context-based Siamese network

- Third, hidden representation $\bar{H}_2$ is input into neural network $\bar{\theta}_2(\cdot)$ with a three-layer fully connected structure. Subsequently, representation $\bar{H}_2$ is transformed into a hidden representation as $\bar{H}_3 = \bar{\theta}_2(\bar{H}_2)$.

After obtaining representation H_3 of the embedding feature extractor and representation $\bar{H}_3$ of the bilateral-context-based feature extractor, the output of the bilateral-context-based Siamese network is calculated using a linear combination function

$$G_w(\phi) = (1-\mu)\cdot H_3 + \mu \cdot \bar{H}_3, \tag{6}$$

where $\phi \in \{S, S\cup t\}$. $\mu \in [0, 1]$ denotes a hyperparameter. The impact of μ on the classifier is discussed in the section "Hyperparameter analysis".

Permutation-invariance-based loss function

To determine whether a new input Chinese entity should be inserted into an existing Chinese synonym set, we use a permutation-invariance-based loss function to train the bilateral-context-based Siamese network classifier. The permutation invariance of the sets is widely used in many fields, such as set expansion, point cloud classification, and outlier detection [43]. For Chinese entity synonym set expansion, permutation invariance is as follows:

Observation 2 *Permutation invariance of entity synonym sets. If a new entity t is synonymous with an existing entity synonym set $S = \{t_1, \ldots, t_n\}$, then $S = \{t_1, \ldots, t_n\}$ and the new set $\bar{S} = \{t_1, \ldots, t_n, t\}$ are permutation invariant in semantics.*

Observation 2 is obvious. For example, the semantics of sets {"The United States", "America"} and {"The United States", "America", "USA"} are identical, because "The United States", "America", and "USA" are synonymous; the semantics of sets {"土豆", "洋芋"} and {"土豆", "洋芋"", "马铃薯"} are identical, where "土豆", "洋芋", and "马铃薯" are the Chinese synonyms for the entity "potato".

Inspired by the above, we minimize the difference between outputs $G_w(S)$ and $G_w(S\cup t)$ to train the bilateral-context-based Siamese network classifier bc-snc(S, t). As depicted in Fig. 3, given entity t and synonym set $S = \{t_1, \ldots, t_n\}$, the difference between $G_w(S)$ and $G_w(S\cup t)$ is calculated, denoted $D_w(S, t) = G_w(S\cup t) - G_w(S)$. Next, we use a sigmoid function to transform $D_w(S, t)$ into score $f_w(S, t)$

$$f_w(S, t) = \text{Sigmoid}(D_w(S, t)). \tag{7}$$

Following Shen et al. [15], we use a log-loss function to train classifier bc-snc(S, t):

$$\mathcal{L} = \log(f_w(S, t))\cdot y - \log(1 - f_w(S, t))\cdot(1-y), \tag{8}$$

where $y = 1$ if entity t can be expanded into synonym set $S = \{t_1, \ldots, t_n\}$ and $y = 0$ otherwise.

Entity synonym set expansion algorithm

This section introduces our Chinese entity synonym set expansion algorithm. In the algorithm, not only the bilateral-context-based Siamese network classifier but also the entity expansion filtering strategy is used for expanding Chinese entity synonym sets.

Entity expansion filtering strategy

According to the definition of the entity synonym set, the synonym entities have the following common feature.

Observation 3 *Domain consistency. Entities belonging to different domains cannot have a synonymous relationship.*

Observation 3 is quite intuitive. For example, on the one hand, entity "番茄 (tomato)" and entity "西红柿 (another Chinese name for tomato)" are synonymous, whereas these two entities and entity "汽车 (automobile)" vary widely. On the other hand, entity "土豆网 (a video website)" and entity "土豆 (potato)" cannot be synonymous, because these two entities belong to different domains.

Based on Observations 1 and 3, we use similarity filtering and domain filtering to filter out wrong synonym entities, thereby mitigating error propagation caused by the Siamese network classifier.

(i) Similarity filtering

Given a new entity t and synonym set $S = \{t_1, \ldots, t_n\}$, $e(t)$ and $e(t_i)$ are embedding representations of entities t and $t_i \in S$. The similarity between t and t_i is calculated as follows:

$$\mathrm{Sim}(t, t_i) = \frac{e(t) \cdot e(t_i)}{\|e(t)\| \cdot \|e(t_i)\|}. \tag{9}$$

Then, the similarity between t and synonym set S is as follows:

$$\overline{\mathrm{Sim}}(S, t) = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathrm{Sim}(t, t_i). \tag{10}$$

(ii) Domain filtering

Similarly, given a new entity t and synonym set $S = \{t_1, \ldots, t_n\}$, the Kullback–Leibler (KL) divergence is used to calculate the domain consistency of t and $t_i \in S$, denoted by

$$\mathrm{KL}(t, t_i) = \sum [p(t) \cdot \log(p(t)) - p(t) \cdot \log(q(t_i))], \tag{11}$$

where $p(t)$ and $q(t_i)$ denote the context distributions for t and t_i, respectively. The domain consistency between t and synonym set S is as follows:

$$\overline{\mathrm{KL}}(S, t) = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathrm{KL}(t, t_i), \tag{12}$$

$$\overline{\mathrm{KL}}\text{-}T(S, t) = 1 - \tanh(\overline{\mathrm{KL}}(S, t)), \tag{13}$$

where $\overline{\mathrm{KL}}\text{-}T(S, t) \in [0, 1]$ is the domain consistency transformed using the tanh function.

Table 1 Dataset statistics

Description	BDSynSetTra	SGSynSetTra
# of entities for training	33,404	4,748
# of synonym sets for training	16,742	2305
# of entities for testing	3861	577
# of synonym sets for testing	1182	255

To balance the effects of similarity filtering and domain filtering on the entity synonym set expansion algorithm, we use a linear function to combine similarity filtering and domain filtering

$$\mathrm{Filter}(S, t) = (1 - \delta) \cdot \overline{\mathrm{Sim}}(S, t) + \delta \cdot \overline{\mathrm{KL}}\text{-}T(S, t), \tag{14}$$

where $\delta \in [0, 1]$ is a hyperparameter. The impact of δ on the algorithm is discussed in the section "Hyperparameter analysis".

Set expansion algorithm

The entity synonym set expansion algorithm is depicted in Algorithm 2. The algorithm uses a bilateral-context-based Siamese network classifier bc-snc(S, t), the Chinese entity vocabulary $V = \{t_1, \ldots, t_i, \ldots, t_{|V|}\}$, an entity expansion filtering score Filter(S, t), and two thresholds κ and λ as input and expands all Chinese entities in V into a Chinese entity synonym set pool $P = \{p_1, \ldots, p_k, \ldots, p_{|N|}\}$.

Particularly, the algorithm traverses all entities $t_i \in V$ to compute the score $f_w(p_k, t_i)$ of classifier bc-snc(S, t) and the filtering score Filter(p_k, t_i) of the entity expansion filtering strategy. If $f_w(p_k, t_i) > \kappa$ and Filter$(p_k, t_i) > \lambda$, then the algorithm adds entity t_i to set p_k. Otherwise, the algorithm expands a new set $p_{|P|+1} = \{t_i\}$ into set pool P, where $|P|$ is the current number of Chinese entity synonym sets in pool P. The entity synonym set expansion algorithm stops after traversing all the entities in the vocabulary.

Experiments

In this section, the proposed approach is applied to two Chinese real-world datasets: BDSynSetTra and SGSynSetTra. Chinese segmentation and part-of-speech (POS) tagging are performed using Hanlp. We employ Word2Vec[4] [41] to train Chinese word embeddings for Baidu Encyclopedia articles and SogouCA. To evaluate the effectiveness of our approach, we compare the results with those of some existing state-of-the-art approaches. Furthermore, we evaluate the impact of

[4] https://radimrehurek.com/gensim/models/word2vec.html.

Algorithm 2 Chinese entity synonym set expansion algorithm

```
Input: (1) A bilateral-context-based siamese network classifier $bc\text{-}snc(S, t)$; (2) a Chinese entity vocabulary $V$; (3) an entity expansion filtering
    score $Filter(S, t)$; (4) two thresholds $\kappa$ and $\lambda$.
Output: A Chinese entity synonym set pool $P = \{p_1, \cdots, p_k, \cdots, p_{|N|}\}$
Initialize the first set $p_1 = \{t_1\}$, add $p_1$ into pool $P = \{p_1\}$;
for each $i \in [2, |V|]$ do //traverse entity vocabulary $V$
    $0 \rightarrow best\_score$; //initialize $best\_score$ to 0
    $1 \rightarrow best\_index$; //initialize $best\_index$ to 1
    for each $k \in [1, |P|]$ do //traverse set pool $P$
        if $f_w(p_k, t_i) > best\_score$ then //measure the score of classifier $bc\text{-}snc(S, t)$
            $f_w(p_k, t_i) \rightarrow best\_score$; //update $best\_score$
            $k \rightarrow best\_index$; //update $best\_index$
        end if
    end for
    if $f_w(p_{best\_index}, t_i) > \kappa$ and $Filter(p_{best\_index}, t_i) > \lambda$ then //evaluate the scores $f_w$ and $Filter$
        $p_{best\_index} \cdot add(t_i)$; //a new entity is added into set $p_{best\_index}$
    else
        $P \cdot append(p_{|P|+1} = \{t_i\})$; //a new set is expanded into pool $P$
    end if
end for
return $P$
```

various values of hyperparameters on the performance of our approach. Finally, we perform ablation and case studies to evaluate the role of bilateral context and entity filtering strategies in improving the performance of the proposed approach.

Experimental settings

Datasets

Table 1 lists the statistics of the BDSynSetTra and SGSynSetTra datasets used for the Chinese entity synonym set expansion task in this study. The descriptions of the two datasets are as follows:

- BDSynSetTra is created from the Baidu Encyclopedia articles. The CN-Dbpedia knowledge base and Hanlp tool are used to process and link the entities in the Baidu Encyclopedia articles. In this dataset, 33,404 entities and 16,742 synonym sets are used for training, and 3861 entities and 1182 synonym sets are used for testing.
- SGSynSetTra is created from the SogouCA corpus. The CN-Dbpedia knowledge base and Hanlp tool are used to process and link the entities in SogouCA. In this dataset, 4748 entities and 2305 synonym sets are used for training, and 577 entities and 255 synonym sets are used for testing.

Benchmark methods for comparison

The following methods are used as benchmarks to compare the performance of the proposed approach.

- **K-means**. K-means[5] clustering algorithm is used to discover Chinese entity synonym entity sets from the Chinese entity vocabulary built from the datasets. We predefine a suitable cluster number K for each dataset. The inputs of the K-means algorithm are the entity embeddings, and the outputs are clustered Chinese entity synonym sets.
- **Birch**. Birch[6] is a hierarchical clustering algorithm. We predefine a suitable cluster number K for each dataset. The inputs of the Birch algorithm are entity embeddings, and the outputs are clustered Chinese synonym entity sets.
- **SVM**. SVM[7] is a supervised approach. First, the approach trains a support vector machine (SVM) classifier to predict Chinese synonym set-instance pairs. Next, the trained SVM classifier is used to expand Chinese entity synonym sets from the datasets.
- **BPNN**. BPNN[8] is a supervised approach. First, the approach trains a back propagation neural network (BPNN) classifier to predict Chinese synonym set-instance pairs. Next, the trained BPNN classifier is used to expand Chinese entity synonym sets from the datasets.
- **SynSetMine**. SynSetMine[9] [15] is a supervised approach. First, the approach trains a Chinese set-instance classifier within embedding and post-transformers to predict Chinese synonym set-instance pairs. Next, the approach uses a set generation algorithm to expand Chinese entity

[5] https://scikit-learn.org/stable/modules/clustering.html#k-means.

[6] https://scikit-learn.org/stable/modules/clustering.html#birch.

[7] https://scikit-learn.org/stable/modules/svm.html.

[8] https://scikit-learn.org/stable/modules/neural_networks_supervised.html.

[9] https://github.com/mickeystroller/SynSetMine-pytorch.

synonym sets from the entity vocabulary built from the datasets.

- **AutoECES**. AutoECES [32] is a supervised approach. First, the approach trains a triplet network classifier to predict Chinese synonym set-instance pairs. Next, the trained triplet network classifier is used to expand Chinese entity synonym sets from the datasets.
- **SynonymNet**. SynonymNet[10] [25] is a supervised approach. First, the approach trains a SynonymNet classifier to predict Chinese synonym set-instance pairs. Next, the trained SynonymNet classifier is used to expand Chinese entity synonym sets from the datasets.
- **CNSynSetE**. CNSynSetE is our proposed approach. In this approach, a bilateral-context-based Siamese network classifier is first designed to predict Chinese synonym set-instance pairs. Next, the approach uses an expansion algorithm within the entity expansion filtering strategy to expand Chinese entity synonym sets from the datasets.

Parameter settings

For fairness of experimental evaluation, all the compared approaches use 100-dimensional Chinese entity embeddings trained using Word2Vec. For embedding and bilateral-context-based feature extractors, the sizes of the two-layer fully connected neural network are 100 and 250, and the sizes of the three-layer fully connected neural network are 250, 500, and 250. The Adam optimizer is used to optimize the bilateral-context-based Siamese network classifier.

In CNSynSetE, there are four hyperparameters, namely, μ, δ, κ, and λ. In particular, μ is the adjustment parameter for bilateral context features, δ is the adjustment parameter for the similarity and domain filtering, κ is a threshold for the bilateral-context-based Siamese network classifier, λ is a threshold for the entity expansion filtering strategy. Particle swarm optimization [44, 45] is employed to obtain the optimal hyperparameter values (see the section "Hyperparameter analysis" for the analysis of parameters). The optimal values of μ, δ, κ, and λ are listed in Table 2.

Metrics

Three common clustering metrics, namely, the Fowlkes–Mallows score[11] (FMI), adjusted Rand index[12] (ARI), and normalized mutual information[13] (NMI), are used to measure the performances of the proposed approach and the selected benchmark approaches.

Table 2 Hyperparameter settings

Parameter	BDSynSetTra	SGSynSetTra
Learning rate	0.00005	0.00005
Dropout rate	0.4	0.4
Negative sample size k	50	50
μ	0.3	0.2
δ	0.7	0.1
κ	0.7	0.1
λ	0.2	0.6

- **FMI**. FMI is usually used to compute the similarity between two given clusters. It is calculated as follows:

$$\mathrm{FMI} = \frac{\mathrm{TP}}{\sqrt{(\mathrm{FP}+\mathrm{TP})\cdot(\mathrm{FN}+\mathrm{TP})}}, \tag{15}$$

  where TP denotes the number of true-positive element pairs belonging to identical clusters in both true labels and prediction labels. FP denotes the number of false-positive element pairs belonging to identical clusters in true labels but not in prediction labels. FN denotes the number of false-negative element pairs belonging to identical clusters in prediction labels but not in true labels.
- **ARI**. ARI is another similarity metric computed using Rand index (RI). It is calculated as follows:

$$\mathrm{RI} = \frac{\mathrm{TP}-\mathrm{TN}}{N}, \tag{16}$$

$$\mathrm{ARI} = \frac{\mathrm{RI}-E(\mathrm{RI})}{\max(\mathrm{RI})-E(\mathrm{RI})}, \tag{17}$$

  where TN denotes the number of true-negative element pairs belonging to identical clusters in both false labels and prediction labels. N is the total number of element pairs.
- **NMI**. NMI is computed using mutual information (MI) and information entropy (IE). It is calculated as follows:

$$\mathrm{NMI}(A,B) = \frac{I(A,B)}{\sqrt{H(A)\cdot H(B)}}, \tag{18}$$

  where $H(A)$ is the IE of A and $I(A,B)$ is the MI between A and B.

In addition, we use the precision (P), recall (R), and F1 scores (F1) to measure the effectiveness of the bilateral-context-based Siamese network classifier in predicting synonym set-instance pairs. The area under the curve (AUC) and mean average precision (MAP) are used to measure the

[10] https://github.com/czhang99/SynonymNet.

[11] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.fowlkes_mallows_score.html.

[12] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html.

[13] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.normalized_mutual_info_score.html.

Table 3 Chinese entity synonym set expansion performance comparison

	BDSynSetTra			SGSynSetTra		
Approach	FMI (%)	ARI (%)	NMI (%)	FMI (%)	ARI (%)	NMI (%)
K-means	3.44	0.73	81.50	34.48	28.73	92.44
Birch	4.96	1.44	83.97	44.56	39.93	94.36
SVM	9.10	5.80	84.25	17.79	11.49	84.11
BPNN	37.61	32.79	92.73	51.97	51.51	94.91
SynSetMine	48.82	48.64	96.06	73.23	73.04	97.24
AutoECES	52.96	52.73	96.11	74.07	73.87	97.27
SynonymNet	54.45	54.19	96.34	74.52	74.14	97.36
CNSynSetE	**60.95**	**60.83**	**96.79**	**81.38**	**81.15**	**98.21**

The best values of experiments are highlighted in bold

performance of the bilateral-context-based Siamese network classifier.

Experimental results

Chinese entity synonym set expansion performance analysis

The results of Chinese entity synonym set expansion tasks obtained using the proposed CNSynSetE and selected benchmark approaches are tabulated in Table 3. From the data in the table, it is evident that CNSynSetE outperforms the selected benchmark approaches in terms of the FMI, ARI, and NMI. An exception is that its FMI, ARI, and NMI results for the BDSynSetTra dataset are lower than those for the SGSynSetTra dataset. This is attributed to the fact that the Chinese synonym entity sets in SGSynSetTra are more common than those in BDSynSetTra. The bilateral-context knowledge, similarity information, and domain information of common Chinese synonym entities are rich and capable of capturing more synonymous semantics to improve the performance of the Chinese entity synonym set expansion task.

From the data in Table 3, the K-means and Birch approaches achieve lower FMI, ARI, and NMI results on the BDSynSetTra and SGSynSetTra datasets, which means that a more elaborate learning model is required to improve the performance of Chinese entity synonym set expansion. In the supervised approach, SVM achieves the lowest FMI, ARI, and NMI results. One possible reason is that the SVM cannot capture enough synonymous features to predict synonym set-instance pairs. The FMI, ARI, and NMI results of BPNN are lower than those of SynSetMine, AutoECES, and SynonymNet. This again indicates that a more elaborate neural network model is required for the Chinese entity synonym set expansion task. Compared with SynSetMine, AutoECES, and SynonymNet, the proposed CNSynSetE approach performs better in terms of the FMI, ARI, and NMI. The aforementioned analysis indicates that the bilateral-context-based Siamese network classifier and entity expansion filtering strategies can improve the performance of the Chinese entity synonym set expansion task.

Table 4 Chinese entity synonym set-pair-based expansion performance comparison

	BDSynSetTra			SGSynSetTra		
Approach	P (%)	R (%)	$F1$ (%)	P (%)	R (%)	$F1$ (%)
SVM	3.33	24.86	5.88	6.76	46.80	11.82
BPNN	22.07	64.09	32.84	58.55	46.13	51.60
SynSetMine	45.05	52.90	48.66	68.96	77.78	73.10
AutoECES	50.34	47.35	49.26	65.34	79.37	71.67
SynonymNet	**59.71**	49.65	54.21	67.88	**81.82**	74.20
CNSynSetE	57.46	**64.66**	**60.85**	**87.21**	75.95	**81.19**

The best values of experiments are highlighted in bold

Table 4 lists the P, R, and $F1$ results of the proposed and selected benchmark approaches for Chinese entity synonym set-pair-based expansion. From the data in the table, it is evident that CNSynSetE outperforms the other approaches in respect of most of the parameters. SVM still achieves the lowest P, R, and $F1$ results. This is because SVM is based on pairwise similarity and does not have a holistic synonymous semantics view of Chinese entity synonym set. An exception is that the P result of SynonymNet on the BDSynSetTra dataset and the R result of SynonymNet on the SGSynSetTra dataset are higher than those of CNSynSetE. However, the $F1$ results of CNSynSetE are higher than those of all of the compared approaches. This again demonstrates that the bilateral-context-based Siamese network classifier and the entity expansion filtering strategies are effective for the Chinese entity synonym set expansion task.

Chinese entity synonym set-instance classifier performance analysis

In this section, we evaluate the performance of the bilateral-context-based Siamese network classifier in Chinese entity

Table 5 Chinese entity synonym set-instance prediction performance comparison

	BDSynSetTra			SGSynSetTra		
Approach	*P* (%)	*R* (%)	*F*1 (%)	*P* (%)	*R* (%)	*F*1 (%)
SVM	9.64	87.31	17.36	79.96	75.61	77.72
BPNN	98.10	87.46	92.48	87.50	34.15	49.12
SynSetMine	96.99	80.46	87.95	**98.16**	79.92	88.11
AutoECES	95.97	80.34	87.46	97.34	80.16	87.92
SynonymNet	98.44	84.08	90.70	97.91	87.80	92.58
CNSynSetE	**98.70**	**88.60**	**93.38**	98.12	**87.99**	**92.78**

The best values of experiments are highlighted in bold

Table 6 Comparison of AUC and MAP for Chinese entity synonym set-instance prediction

	BDSynSetTra		SGSynSetTra	
Approach	AUC (%)	MAP (%)	AUC (%)	MAP (%)
SVM	61.44	57.60	89.42	81.37
BPNN	99.71	99.09	94.62	85.99
SynSetMine	99.78	98.79	99.72	98.91
AutoECES	99.38	97.82	99.06	97.40
SynonymNet	99.85	99.29	99.89	99.35
CNSynSetE	**99.91**	**99.61**	**99.92**	**99.60**

The best values of experiments are highlighted in bold

synonym set-instance prediction. Table 5 lists the *P*, *R*, and *F*1 results on Chinese entity synonym set-instance prediction obtained using various approaches. SVM still achieves the lowest *P*, *R*, and *F*1 results on the BDSynSetTra and SGSynSetTra datasets. The *P*, *R*, and *F*1 results of BPNN on the BDSynSetTra dataset are higher than those on the SGSynSetTra dataset. Compared with SynSetMine, AutoECES, and SynonymNet, CNSynSetE obtains the highest *F*1 results on the BDSynSetTra and SGSynSetTra datasets. These results indicate that the Siamese network classifier combined with bilateral context can effectively improve the performance of Chinese entity synonym set-instance prediction.

A comparison of precision–recall curves of all approaches is depicted in Fig. 7. In general, it is evident that CNSynSetE achieves higher precision results when compared with the other approaches for the whole range of recall results. An exception is that SVM performs significantly worse when compared with the other approaches. For BPNN, the results of precision–recall curves on the BDSynSetTra dataset are higher than those on the SGSynSetTra dataset. In addition, we found little difference in the results of the precision–recall curves for CNSynSetE, SynonymNet, AutoECES, and SynSetMine. However, the FMI and ARI results (see Table 3) of SynonymNet, AutoECES, and SynSetMine are lower than those of CNSynSetE. This again implies that the bilateral-context-based Siamese network classifier and entity expansion filtering strategies can effectively improve the performance of the Chinese entity synonym set expansion task.

Table 6 lists the AUC and MAP results for Chinese entity synonym set-instance prediction obtained using various approaches. The AUC and MAP results of SVM are still lower than those of the other approaches. For BPNN, the AUC and MAP results on the BDSynSetTra dataset are higher than those on the SGSynSetTra dataset. Similar to the precision–recall curves, the AUC and MAP results of

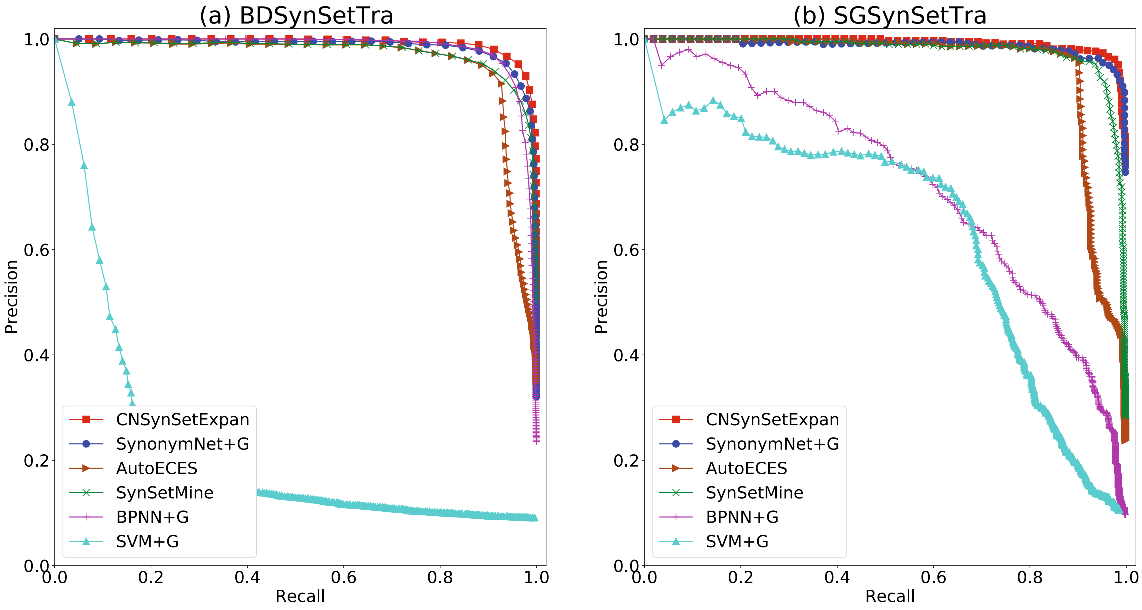


Fig. 7 Comparison of precision–recall curves for Chinese entity synonym set-instance prediction

CNSynSetE, SynonymNet, AutoECES, and SynSetMine are not very different. This means that relying on an excellent prediction model alone may not achieve good results in the two-step Chinese entity synonym set expansion task.

Hyperparameter analysis

In the aforementioned results, we used the optimal hyperparameters μ, δ, κ, and λ (see Table 2) to analyze the performance of CNSynSetE. To further analyze the effects of these hyperparameters on CNSynSetE, we conduct a detailed performance analysis for each hyperparameter. The performance results for various values of these hyperparameters are depicted in Fig. 8.

- μ **analysis**. μ is the adjustment parameter for bilateral context features. We fix hyperparameters δ, κ, and λ as the optimal parameter values and assign a value between 0.1 and 0.9 to hyperparameter μ. In Fig. 8a, the FMI values of CNSynSetE are stable on the BDSynSetTra and SGSynSetTra datasets. In particular, on the BDSynSetTra dataset, CNSynSetE obtains a higher FMI value when $\mu = 0.3$; on the SGSynSetTra dataset, CNSynSetE obtains a higher FMI value when $\mu = 0.2$.
- δ **analysis**. δ is the adjustment parameter for similarity and domain filtering. We fix hyperparameters μ, κ, and λ as the optimal parameter values and assign a value between 0.1 and 0.9 to hyperparameter δ. It is evident from Fig. 8b that the FMI values of CNSynSetE decrease with a increase in δ for the SGSynSetTra dataset. The FMI values of CNSynSetE first increase and then decrease with an increase in δ for the BDSynSetTra dataset. On the BDSynSetTra dataset, CNSynSetE obtains a higher FMI value when $\delta = 0.7$; on the SGSynSetTra dataset, CNSynSetE obtains a higher FMI value when $\delta = 0.1$.
- κ **analysis**. κ is the threshold for the bilateral-context-based Siamese network classifier. We fix hyperparameters μ, δ, and λ as the optimal parameter values and assign a value between 0.1 and 0.9 to hyperparameter κ. It is evident from Fig. 8c that the FMI values of CNSynSetE decrease with an increase in κ for the SGSynSetTra dataset. The FMI values of CNSynSetE first increase and then decrease with an increase in κ for the BDSynSetTra dataset. On the BDSynSetTra dataset, CNSynSetE obtains a higher FMI value when $\kappa = 0.7$; on the SGSynSetTra dataset, CNSynSetE obtains a higher FMI value when $\kappa = 0.1$.
- λ **analysis**. λ is the threshold for the entity expansion filtering strategy. We fix the hyperparameters μ, δ, and κ as

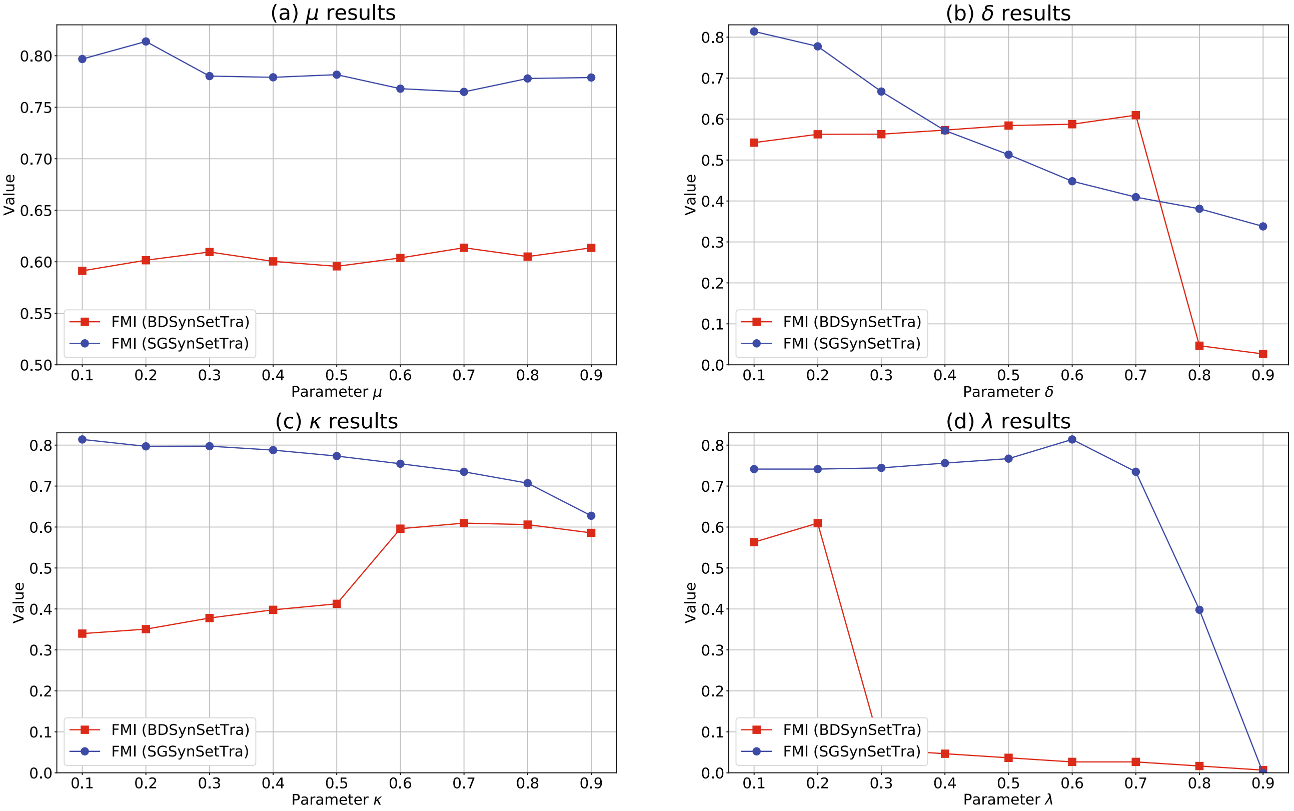


Fig. 8 Performance results for various values of hyperparameters

Table 7 Comparison of time consumption

Approach	Training		Prediction	
	BDSynSetTra	SGSynSetTra	BDSynSetTra	SGSynSetTra
K-means	–	–	8.86 s	1.55 s
Bitch	–	–	9.74 s	2.16 s
SVM	1.2 h	6 m	6.46 s	1.38 s
BPNN	8 h	49 m	5.12 s	1.14 s
SynSetMine	9.1 h	51 m	5.46 s	1.21 s
AutoECES	9.3 h	57 m	7.28 s	1.57 s
SynonymNet	10.2 h	1.1 h	7.17 s	1.45 s
CNSynSetE	9.4 h	53 m	6.29 s	1.28 s

the optimal parameter values and assign a value between 0.1 and 0.9 to hyperparameter λ. It is evident from Fig. 8d that both the FMI values of CNSynSetE first increase and then decrease with an increase in λ for the BDSynSetTra and SGSynSetTra datasets. On the BDSynSetTra dataset, CNSynSetE obtains a higher FMI value when $\lambda = 0.2$; on the SGSynSetTra dataset, CNSynSetE obtains a higher FMI value when $\lambda = 0.6$.

Based on these analyses, setting $\mu = 0.3$, $\delta = 0.7$, $\kappa = 0.7$, and $\lambda = 0.2$ is recommended for the BDSynSetTra dataset, and setting $\mu = 0.2$, $\delta = 0.1$, $\kappa = 0.1$, and $\lambda = 0.6$ is recommended for the SGSynSetTra dataset.

Effect of negative sample size

To evaluate the impact of negative sample size k on the bilateral-context-based Siamese network classifier, this section evaluates how different negative sample sizes will affect the performance of bilateral-context-based Siamese network classifier on BDSynSetTra and SGSynSetTra datasets.

The experimental results of different negative sample sizes are shown in Fig. 9. We find that both the FMI values increase with an increase in negative sample size k for the BDSynSetTra and SGSynSetTra datasets. Thus, the value of negative sample size k in the range of 30–60 is recommended for the BDSynSetTra dataset, and the value of negative sample size k in the range of 20–60 is recommended for the SGSynSetTra dataset.

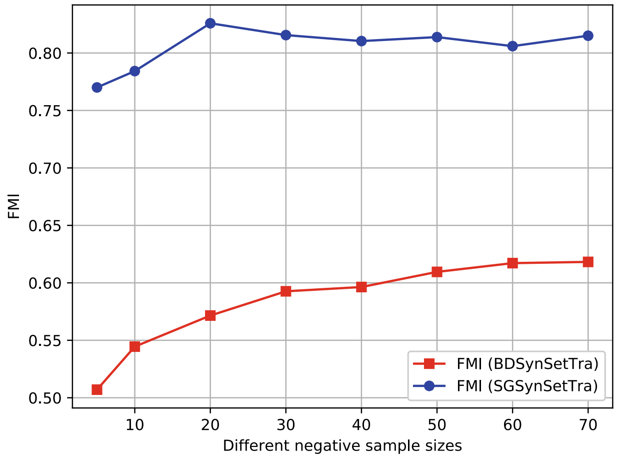


Fig. 9 Performance results for different negative sample sizes

Time consumption

This section gives the time consumption of the comparison approaches. The PyTorch library is used to implement the neural network models (BPNN, SynSetMine, AutoECES, SynonymNet, and CNSynSetE). The clustering models (K-means and Bitch) and SVM are run on CPU and the neural network models are run on Quadro RTX6000 GPU.

Comparison results are listed in Table 7. The time consumption of the proposed CNSynSetE is close to the other models. In the neural network models, BPNN obtains the faster prediction, but its performance is low. SynSetMine is faster than AutoECES and SynonymNet. CNSynSetE is relatively slower than SynSetMine. This is because CNSynSetE integrates bilateral contexts into the Siamese network, which sacrifices a little time to process the bilateral contexts. However, considering the time consumption and the aforementioned metrics, CNSynSetE is an effective approach for discovering Chinese entity synonym sets.

Ablation study

To further analyze the impact of the subcomponents of our proposed approach (e.g., bilateral context information, similarity filtering strategy, and domain filtering strategy) on its overall performance, we divide the proposed approach into four ablation approaches: No-BiContext, No-FiltStrategy, No-SimFiltering, and No-DomFiltering.

- **No-BiContext**. No-BiContext is an ablation approach that does not use bilateral context information. First, this strategy uses a Siamese network classifier to predict synonym set-instance pairs. Next, it uses an expansion

Table 8 Performance comparison for our approach, compared approach, and ablation approaches to Chinese entity synonym set expansion

Approach	BDSynSetTra			SGSynSetTra		
	FMI (%)	ARI (%)	NMI (%)	FMI (%)	ARI (%)	NMI (%)
No-BiContext	51.74	51.55	96.18	77.44	77.39	97.67
No-FiltStrategy	56.22	55.65	96.18	74.15	73.38	97.34
No-SimFiltering	2.66	0.14	94.94	32.82	23.01	94.34
No-DomFiltering	56.25	55.68	96.19	78.24	77.73	97.89
SynSetMine	48.82	48.64	96.06	73.23	73.04	97.24
AutoECES	52.96	52.73	96.11	74.07	73.87	97.27
SynonymNet	54.45	54.19	96.34	74.52	74.14	97.36
CNSynSetE	**60.95**	**60.83**	**96.79**	**81.38**	**81.15**	**98.21**

The best values of experiments are highlighted in bold

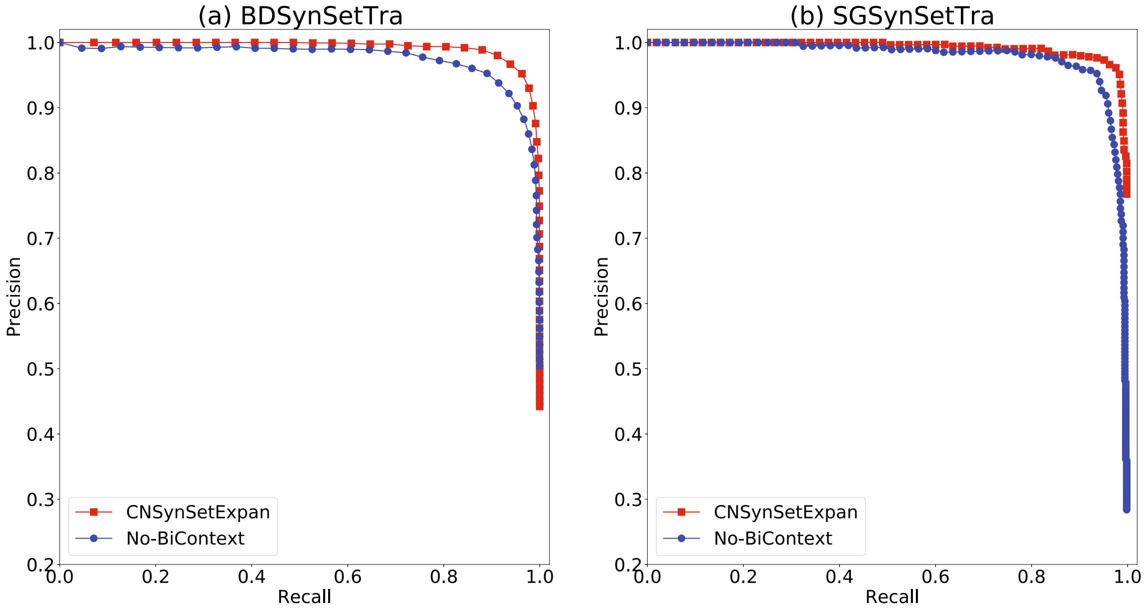


Fig. 10 Precision–recall curves for CNSynSetE and No-BiContext approaches

algorithm with similarity and domain filtering strategies to expand Chinese entity synonym sets.

- **No-FiltStrategy**. No-FiltStrategy is an ablation approach that does not use any filtering strategy. First, this strategy uses a bilateral-context-based Siamese network classifier to predict synonym set-instance pairs. Next, it uses an expansion algorithm without using the similarity and domain filtering strategies to expand Chinese entity synonym sets.
- **No-SimFiltering**. No-SimFiltering is an ablation approach that does not use the similarity filtering strategy. First, this approach uses a bilateral-context-based Siamese network classifier to predict synonym set-instance pairs. Next, it uses an expansion algorithm with only a domain filtering strategy to expand Chinese entity synonym sets.
- **No-DomFiltering**. No-DomFiltering is an ablation approach that does not use the domain filtering strategy. First, this approach uses a bilateral-context-based Siamese network classifier to predict synonym set-instance pairs. Next, it uses an expansion algorithm with only a similarity filtering strategy to expand Chinese entity synonym sets.

Table 8 lists the experimental results of CNSynSetE, SynSetMine, AutoECES, SynonymNet, and the aforementioned ablation approaches to Chinese entity synonym set expansion. It is evident from the data in the table that CNSynSetE achieves the best experimental results in terms of FMI, ARI, and NMI. An exception is that No-SimFiltering achieves the lowest experimental results, meaning that the Chinese entity synonym set expansion algorithm with only a domain filtering strategy worsens the performance of CNSynSetE. The experimental results of No-FiltStrategy are close to SynSetMine, AutoECES, and SynonymNet. However, the experimental results of CNSynSetE are better than SynSetMine, AutoECES, and SynonymNet when adding the

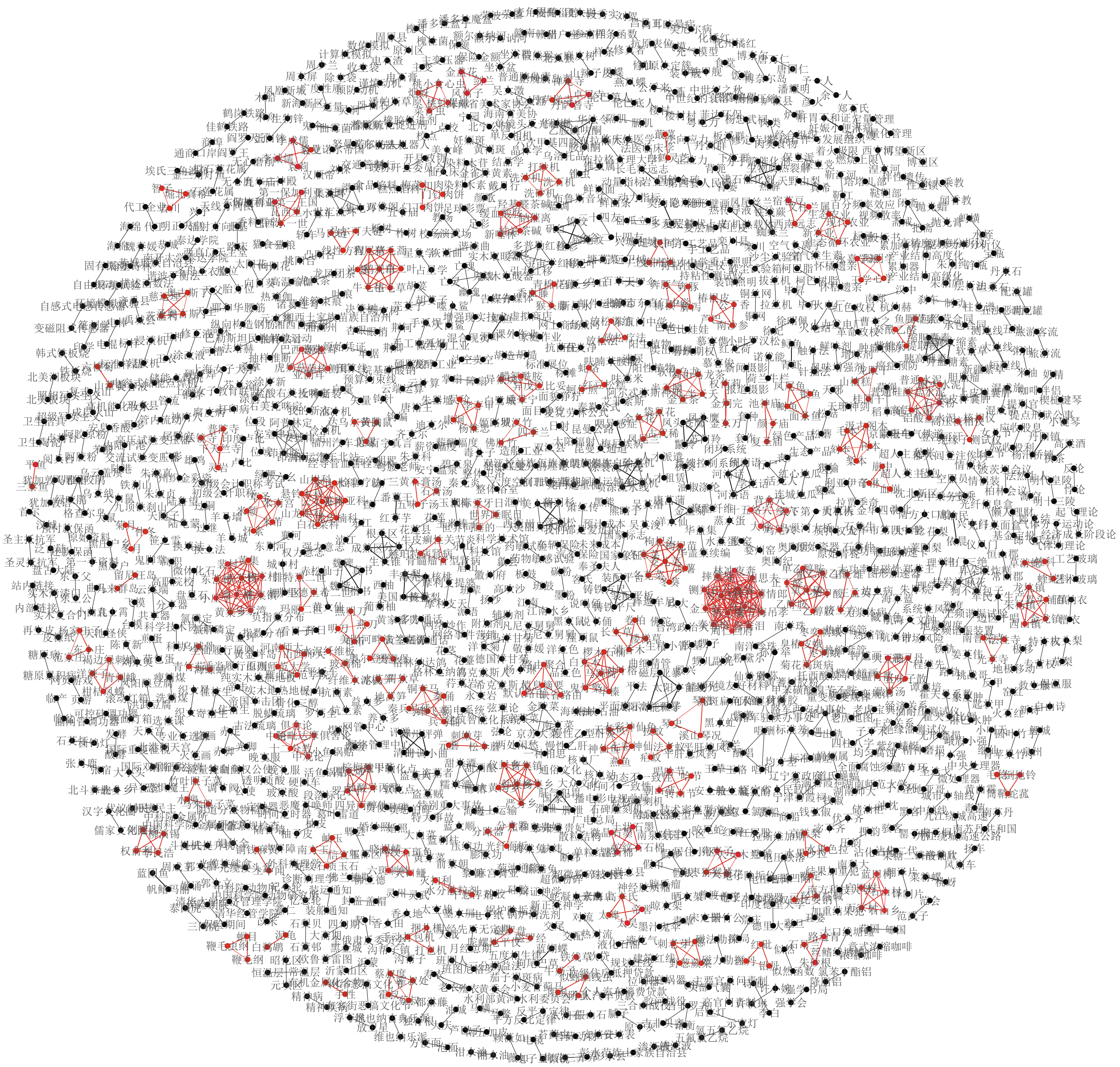


Fig. 11 Output Chinese entity synonym sets obtained using the proposed approach on the BDSynSetTra dataset. Black vertices and edges denote the correct output synonym sets. Red vertices and edges denote the wrong output synonym sets

filtering strategies (similarity filtering and domain filtering). The above analysis indicates that similarity filtering and domain filtering play a positive role in improving the performance of CNSynSetE in the Chinese entity synonym set expansion task.

To further evaluate the impact of bilateral context information, we compare the bilateral-context-based Siamese network classifier of CNSynSetE with the Siamese network classifier of No-BiContext. Figure 10 depicts the precision–recall curves for CNSynSetE and No-BiContext. The precision results of CNSynSetE are higher than those of No-BiContext considering the whole range of recall results. This proves that integrating the bilateral context information into the Siamese network classifier can indeed improve the performance of the Chinese entity synonym set expansion task.

Case study

To verify the effectiveness of our proposed approach and analyze the reasons for the incorrect results generated by our approach, a case study is presented in this subsection.

Figures 11 and 12 depict the visual output Chinese entity synonym sets obtained using the proposed approach on the BDSynSetTra and SGSynSetTra datasets, respectively. The

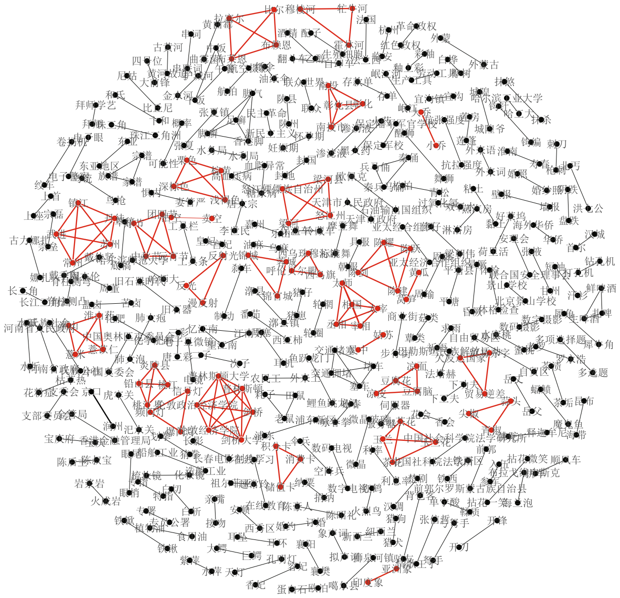


Fig. 12 Output Chinese entity synonym sets obtained using the proposed approach on the SGSynSetTra dataset. Black vertices and edges denote the correct output synonym sets. Red vertices and edges denote the wrong output synonym sets

black vertices and edges denote the correct output synonym sets. The red vertices and edges denote the wrong output synonym sets. It is evident from these figures that most of the Chinese entity synonym sets with sizes 2 and 3 are correct results. However, we report the following error case: when the size of the synonym set continues to increase, the error rate of the set continues to increase. To further analyze the causes for this error case, as listed in Table 9, some Chinese entity synonym set outputs of our approach are randomly selected (only ten synonym sets are given for each dataset owing to space limitations).

It is evident from the data in Table 9 that some selected synonym sets of size 4 or larger are wrong cases. For example, {兵马俑,秦兵马俑,秦俑,马踏飞燕,铜奔马} in BDSynSetTra and {元宵节,中秋节,灯节,团圆节} in SGSynSetTra are wrong cases. The reasons are as follows:

- On the one hand, the semantic information in the Chinese synonym set becomes more complex when the size of the synonym set increases. This prevents the proposed approach from capturing more synonymous information to predict the Chinese entity synonym sets.
- On the other hand, some entities are so similar in semantics that our approach cannot identify whether they are synonymous relations or related-to relations. It is still difficult to discriminate between synonymous relations and related-to relations for Chinese entities because of their size and breadth.

Conclusion

This paper proposes a bilateral context and filtering strategy-based approach to generate Chinese entity synonym sets. Specifically, a bilateral-context-based Siamese network classifier is developed to evaluate an input Chinese entity for its inclusion into the existing synonym set. The classifier is capable of imposing soft holistic semantic constraints to improve synonym prediction. To generate Chinese entity synonym sets, a filtering-strategy-based set expansion algorithm is presented. The filtering strategy are capable of enhancing semantic and domain consistencies to filter out

Table 9 Chinese entity synonym set output examples (O denotes the output results of our approach, and G denotes the ground truth)

BDSynSetTra					
Output	O	G	Output	O	G
{公鸡,雄鸡}	1	1	{第一手资料,原始资料}	1	1
{安搏律定,阿普林定,茚满丙二胺,安室律定,茚丙胺}	1	1	{婚外情,外遇,出轨}	1	1
{向日葵,太阳花,向阳花}	1	1	{银鱼,面条鱼,鲥鱼,凤尾鱼}	1	0
{假钱,假钞,假币}	1	1	{赤脚,光脚,赤足}	1	1
{兵马俑,秦兵马俑,秦俑,马踏飞燕,铜奔马}	1	0	{龙眼干,桂圆干,山楂糕}	1	0
SGSynSetTra					
Output	O	G	Output	O	G
{堵车,交通拥堵,交通堵塞,塞车}	1	1	{脚癣,足癣,香港脚,脚气}	1	1
{午饭,午餐,中饭}	1	1	{服务器,伺服器}	1	1
{哈尔滨工业大学,哈工大}	1	1	{元宵节,中秋节,灯节,团圆节}	1	0
{妊娠期,怀孕期}	1	1	{宗谱,家谱,族谱}	1	1
{储值卡,消费卡,积分卡}	1	0	{猎狗,猎犬}	1	1

wrong Chinese synonym entities and mitigate the problem of error propagation caused by the Siamese network classifier. The proposed approach and some state-of-the-art benchmark approaches are applied to two Chinese real-world synonym set datasets to evaluate their comparative performances. The experimental results indicate that the proposed approach is effective and outperforms the selected state-of-the-art approaches in the Chinese entity synonym set expansion task.

In the future, we intend to expand more Chinese entity synonym sets from other Chinese text corpora (e.g., news text corpora). We also intend to use a multimodal-data-based method to discriminate between synonymous entities and related-to entities and improve the accuracy of the final expanded Chinese entity synonym sets. Furthermore, the use of the proposed approach in combination with Bidirectional Encoder Representations from Transformers (BERT) [46], as an alternative approach for expanding Chinese entity synonym sets, could be explored.

Acknowledgements This research was funded by the University Natural Science Research Projects of Anhui Province under Grant No. 2022AH050972. It was jointly funded by the Excellent Young Talents Fund Program of Higher Education Institutions of Anhui Province under Grant No. gxyq2020031 and the University Natural Science Research Projects of Anhui Province under Grant Nos. KJ2021A0516 and KJ2020A0361.

Data availability The data are available from the corresponding author on reasonable request.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Mahdisoltani F, Biega J, Suchanek FM (2015) YAGO3: a knowledge base from multilingual wikipedias. In: Seventh biennial conference on innovative data systems research, CIDR 2015, Asilomar, CA, USA, January 4–7, 2015
2. Xu B, Xu Y, Liang J, Xie C, Liang B, Cui W, Xiao Y (2017) Cn-dbpedia: a never-ending Chinese knowledge extraction system. In: Advances in artificial intelligence: from theory to practice—30th international conference on industrial engineering and other applications of applied intelligent systems. IEA/AIE 2017, Arras, France, June 27–30, part II, vol 10351, pp 428–438
3. Qi F, Chang L, Sun M, Ouyang S, Liu Z (2020) Towards building a multilingual sememe knowledge base: Predicting sememes for BabelNet synsets. In: The thirty-fourth AAAI conference on artificial intelligence, AAAI 2020, the thirty-second innovative applications of artificial intelligence conference, IAAI 2020, the tenth AAAI symposium on educational advances in artificial intelligence, EAAI 2020, New York, NY, USA, February 7–12, pp 8624–8631
4. Rios-Alvarado AB, Martinez-Rodriguez JL, Garcia-Perez AG, Guerrero-Melendez TY, Lopez-Arevalo I, Gonzalez-Compean JL (2022) Exploiting lexical patterns for knowledge graph construction from unstructured text in Spanish. Complex Intell Syst 9:1281–1297
5. Gupta A, Lebret R, Harkous H, Aberer K (2017) Taxonomy induction using hypernym subsequences. In: Proceedings of the 2017 ACM on conference on information and knowledge management. CIKM 2017, Singapore, November 06–10, pp 1329–1338
6. Huang S, Luo X, Huang J, Guo Y, Gu S (2019) An unsupervised approach for learning a Chinese IS-A taxonomy from an unstructured corpus. Knowl Based Syst 182:104861
7. Huang S, Luo X, Huang J, Wang H, Gu S, Guo Y (2020) Improving taxonomic relation learning via incorporating relation descriptions into word embeddings. Concurr Comput: Pract Exp 32(14):e5696
8. Shen J, Shen Z, Xiong C, Wang C, Wang K, Han J (2020) TaxoExpan: self-supervised taxonomy expansion with position-enhanced graph neural network. In: Huang Y, King I, Liu T, van Steen M (eds) WWW '20: the web conference 2020, Taipei, Taiwan, April 20–24, pp 486–497
9. Gu S, Luo X, Wang H, Huang J, Wei Q, Huang S (2021) Improving answer selection with global features. Expert Syst: J Knowl Eng 38(1):e12603
10. Bakhshi M, Nematbakhsh M, Mohsenzadeh M, Rahmani AM (2022) SParseQA: sequential word reordering and parsing for answering complex natural language questions over knowledge graphs. Knowl Based Syst 235:107626
11. Li X, Alazab M, Li Q, Yu K, Yin Q (2022) Question-aware memory network for multi-hop question answering in human–robot interaction. Complex Intell Syst 8:851–861
12. Shen J, Qiu W, Shang J, Vanni M, Ren X, Han J (2020) Synsetexpan: an iterative framework for joint entity set expansion and synonym discovery. In: Proceedings of the 2020 conference on empirical methods in natural language processing EMNLP 2020, Online, November 16–20, pp 8292–8307
13. Huang S, Luo X, Huang J, Qin W, Gu S (2020a) Neural entity synonym set generation using association information and entity constraint. In: 2020 IEEE international conference on knowledge graph, ICKG 2020, Online, August 9–11, pp 321–328
14. Yang Y, Yin X, Yang H, Fei X, Peng H, Zhou K, Lai K, Shen J (2021) KGSynNet: a novel entity synonyms discovery framework with knowledge graph. In: Database systems for advanced applications—26th international conference. DASFAA 2021, Taipei, China, April 11–14, part I, vol 12681, pp 174–190
15. Shen J, Lyu R, Ren X, Vanni M, Sadler BM, Han J (2019) Mining entity synonyms with efficient neural set generation. In: The thirty-third AAAI conference on artificial intelligence, AAAI 2019, the thirty-first innovative applications of artificial intelligence conference, IAAI 2019, the ninth AAAI symposium on educational advances in artificial intelligence, EAAI 2019, Honolulu, HI, USA, January 27–February 1, pp 249–256
16. McCrae JP, Collier N (2008) Synonym set extraction from the biomedical literature by lexical pattern discovery. BMC Bioinform 9:159
17. Wang W, Thomas C, Sheth AP, Chan V (2010) Pattern-based synonym and antonym extraction. In: Proceedings of the 48th annual

southeast regional conference, Oxford, MS, USA, April 15–17, p 64

18. Li W, Lu Q (2011) A hybrid extraction model for Chinese noun/verb synonymous bi-gram collocations. In: Proceedings of the 25th Pacific Asia conference on language, information and computation, PACLIC 25, Singapore, December 16–18, pp 430–439
19. Nguyen KA, im Walde SS, Vu NT (2017) Distinguishing antonyms and synonyms in a pattern-based neural network. In: Proceedings of the 15th conference of the European chapter of the association for computational linguistics. EACL 2017, Valencia, Spain, April 3–7, pp 76–85
20. Harris ZS (1954) Distributional structure. Word 10(2–3):146–162
21. Qu M, Ren X, Han J (2017) Automatic synonym discovery with knowledge bases. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, Halifax, NS, Canada, August 13–17, pp 997–1005
22. Turney PD (2001) Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In: Machine learning: EMCL 2001, 12th European conference on machine learning, Freiburg, Germany, September 5–7, vol 2167, pp 491–502
23. Chakrabarti K, Chaudhuri S, Cheng T, Xin D (2012) A framework for robust discovery of entity synonyms. In: The 18th ACM SIGKDD international conference on knowledge discovery and data mining. KDD'12, Beijing, China, August 12–16, pp 1384–1392
24. Ma X, Luo X, Huang S, Guo Y (2019) Multi-distribution characteristics based Chinese entity synonym extraction from the web. Int J Intell Inf Technol 15(3):42–63
25. Zhang C, Li Y, Du N, Fan W, Yu PS (2020) Entity synonym discovery via multipiece bilateral context matching. In Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI 2020, pp 1431–1437
26. Dorow B, Widdows D (2003) Discovering corpus-specific word senses. In: EACL 2003, 10th conference of the European chapter of the association for computational linguistics, Budapest, Hungary, April 12–17, pp 79–82
27. Duan L, Chen J, Li H, Li A (2010) A Chinese synonyms reduced algorithm based on sememe tree. In: International conference on computational aspects of social networks. CASON 2010, Taiyuan, China, September 26–28, pp 337–340
28. Ustalov D, Panchenko A, Biemann C (2017) Automatic induction of synsets from a graph of synonyms. In: Proceedings of the 55th annual meeting of the association for computational linguistics, ACL 2017, Vancouver, Canada, July 30–August 4, vol 1, Long papers, pp 1579–1590
29. Ercan G, Haziyev F (2019) Synset expansion on translation graph for automatic wordnet construction. Inf Process Manag 56(1):130–150
30. Ren X, Cheng T (2015) Synonym discovery for structured entities on heterogeneous graphs. In: Proceedings of the 24th international conference on world wide web companion. WWW 2015, Florence, Italy, May 18–22, pp 443–453
31. Shen J, Wu Z, Lei D, Shang J, Ren X, Han J (2017) Setexpan: corpus-based set expansion via context feature selection and rank ensemble. In: Machine learning and knowledge discovery in databases—European conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, vol 10534, pp 288–304
32. Huang S, Qin W, Zhao S, Gu S (2020c) An automatic approach for extracting Chinese entity synonyms from encyclopedias. In: Proceedings of the 2020 3rd international conference on big data technologies, ICBDT, Qingdao, China, September 18–20
33. Wang C, Yan J, Zhou A, He X (2017) Transductive non-linear learning for Chinese hypernym prediction. In: Proceedings of the 55th annual meeting of the association for computational linguistics. ACL 2017, Vancouver, Canada, July 30–August 4, vol 1, Long papers, pp 1394–1404
34. Hearst MA (1992) Automatic acquisition of hyponyms from large text corpora. In: The 14th international conference on computational linguistics. COLING 1992, Nantes, France, August 23–28, pp 539–545
35. Kwong OY, Tsou BK (2006) Feasibility of enriching a Chinese synonym dictionary with a synchronous Chinese corpus. In: Advances in natural language processing, 5th international conference on NLP. FINTAL 2006, Turku, Finland, August 23–25, vol 4139, pp 322–332
36. Yu L-C, Chien W-N, Chen S-T (2011) A baseline system for Chinese near-synonym choice. In: Fifth international joint conference on natural language processing. IJCNLP 2011, Chiang Mai, Thailand, November 8–13, pp 1366–1370
37. Gan Y (2017) A study on Chinese synonyms: from the perspective of collocations. In: Chinese lexical semantics—18th workshop. CLSW 2017, Leshan, China, May 18–20, revised selected papers, vol 10709, pp 586–600
38. Lu Y, Hou H (2008) Research on automatic acquiring of Chinese synonyms from wiki repository. In: Proceedings of the 2008 IEEE/WIC/ACM international conference on web intelligence and international conference on intelligent agent technology—workshops. Sydney, NSW, Australia, December 9–12, pp 287–290
39. Mintz M, Bills S, Snow R, Jurafsky D (2009) Distant supervision for relation extraction without labeled data. In: ACL 2009: proceedings of the 47th annual meeting of the association for computational linguistics and the 4th international joint conference on natural language processing of the AFNLP. Singapore, August 2–7, pp 1003–1011
40. Vashishth S, Joshi R, Prayaga SS, Bhattacharyya C, Talukdar PP (2018) RESIDE: improving distantly-supervised neural relation extraction using side information. In: Proceedings of the 2018 conference on empirical methods in natural language processing, Brussels, Belgium, October 31–November 4, pp 1257–1266
41. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. In: 1st International conference on learning representations, ICLR 2013, Scottsdale, AZ, USA, May 2–4 (workshop track proceedings)
42. Ji G, Liu K, He S, Zhao J (2017) Distant supervision for relation extraction with sentence-level attention and entity descriptions. In: Proceedings of the 31st AAAI conference on artificial intelligence, San Francisco, CA, USA, February 4–9, pp 3060–3066
43. Zaheer M, Kottur S, Ravanbakhsh S, Poczos B, Salakhutdinov R, Smola AJ (2017) Deep sets. In: Advances in neural information processing systems 30: annual conference on neural information processing systems 2017, Long Beach, CA, USA, December 4–9, pp 3391–3401
44. Kennedy J, Eberhart R (1995) Particle swarm optimization. In: Proceedings of international conference on neural networks (ICNN–95), Perth, WA, Australia, November 27–December 1, pp 1942–1948
45. Wang Z-J, Zhan Z-H, Yu W, Lin Y, Zhang J, Gu T, Zhang J (2020) Dynamic group learning distributed particle swarm optimization for large-scale optimization and its application in cloud workflow scheduling. IEEE Trans Cybern 50(6):2715–2729
46. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies. NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, vol 1 (long and short papers), pp 4171–4186