Front.Comput.Sci.
DOI

LETTER

Supplementary Material of "Integrating Heterogeneous Thesauruses for Chinese Synonyms"

Jianbing ZHANG, Peng WU, Yingjie ZHANG, Shujian HUANG*, **Xinyu DAI, Jiajun CHEN**

National Key Laboratory for Novel Software Technology,
Nanjing University, Nanjing 210023, China

Abstract Linguistic resource plays a crucial role in computing linguistics. Different resources may have different structures due to their proposed purposes and organizations. However, it is difficult to use these heterogeneous resources in one model. In this paper, we propose an effective yet simple method which can integrate the heterogeneous synonyms in Chinese by hierarchical mapping. The proposed method can combine the clear structures and Chinese-specific lexemes which are the advantages of different Chinese thesauruses. With the development of deep learning, word embedding can represent word meaning into a dense vector, and in this paper we use word2vec to evaluate our model. Experimental results show that the proposed method can achieve more synonym sets, and the meaning in the synonym set is more concentrated.

Keywords heterogeneous thesauruses, Chinese synonyms

1 Introduction

Lexical semantic resource plays an important role in natural language processing. So far, many lexical semantic resources have been developed by the world-wide linguists, such as WordNet [1], ConceptNet [2,3] in English, HowNet [4], Chinese Concept Dictionary (CCD) [5], and Tongyici-Cilin (Cilin) [6] in Chinese. Different recourses usually have different focuses and structures, while some of them are also closely related and could be complementary to each other. As a result, the integration of several resources may be more useful than only using one of them for a certain purpose.

There are roughly two types of methods to integrate two different lexical semantic resources. The first one is in semantic description level, and the other is in ontology level. For the former type, Carpuat [7] links synonym sets (Synsets) in WordNet to definitions (DEFs), which is described via a set of pre-defined basic concepts called "semantic atoms", based on the English-Chinese bilingual corpus and dictionaries. Mei [8] investigates a method that integrates Cilin and HowNet. This method uses DEFs in HowNet to describe the word sets in Cilin, and tags these DEFs by word sets IDs as well. For the latter type, a typical work is performed by Niles and Pease [9], which maps the WordNet Synsets onto the ontologies in the suggested upper merged ontology (SUMO) and achieves a novel computable lexical resources in English. However, it is hard to find such existing works that focus on the Chinese resources.

In this paper, we propose an effective yet simple method to integrate two heterogeneous thesauruses, i.e., CCD and Cilin, to build a large resource for Chinese synonyms. By integrating these dictionaries, the proposed method can benefit from both the clear and rich semantic relations of CCD and the large set of Chinese-specific lexemes in Cilin. With the development of deep learning in natural language processing, word embedding can represent word meaning as a dense vector [10,11], and in this paper we use word2vec to evaluate our model. Experimental results show that the proposed method can

Received July 12, 2019; accepted November 04, 2019

E-mail: huangsj@nju.edu.cn

Table 1: Statistics of CCD Categories and Cilin Classes for each POS

POS	Class	Cilin Medium Class	Subdivision	CCD Category
Noun	A B C D	48 –	588	26
Verb	F G H I J	34	567	15
Adjective	E	6 –	179	3
Adverb	K	1	35	1
Pronoun	A B C E	5 –	8	0
Others	K L	6	48	0

achieve more synonym sets, and the meaning in the synonym set is more concentrated.

The rest of this paper is organized as follows. In Section 2, we compare the difference between CCD and Cilin. In Section 3 we show a simple method which integrates two resources directly. And in Section 4, we propose a hierarchical mapping procedure for the integration. Section 5 evaluates the mapping results, and finally Section 6 gives the conclusion.

2 Comparison between CCD and Cilin

In this section, we compares CCD [5] with Cilin [6] from the aspects of vocabulary, classification system, and organization, respectively.

2.1 Vocabulary

CCD [5] is a Chinese thesaurus on the basis of WordNet [1]. It collects 125,929 words (or phrases) into 99,642 semantic classes (called CSynsets), including nouns, verbs, adjectives and adverbs. All CSynsets in CCD show the common concepts between Chinese and English, but miss a large number of Chinese-specific words.

In contrast to CCD, Cilin [6] is a thesaurus with a wide range of Chinese-specific phenomena, including words, morphemes, phrases and idioms, even some common dialects or ancient words. Both content and function words are collected. There are 77,457 words in Cilin, with only about half of them (35,123 words) covered by CCD. However, Cilin may not collect all senses of a word, because its vocabulary is focused on the synonyms.

As a result of the different purposes of CCD and Cilin, two "monosemous" words with the same shape in them respectively may convey totally different meanings.

2.2 Classification System

For the taxonomy of CCD, it follows WordNet and divides all CSynsets into 45 categories. In CCD, the CSynsets in the same category are under the same part-of-speech (POS).

By contrast, all the concepts in Cilin are separated into 12 classes, 94 medium classes (MC), and 1,425 subdivisions. For content words, there are 4 classes of nouns, 5 classes of verbs, 1 class of adjectives and 1 class of adverbs. While for function words, they are scattered in all classes. More details are shown in Table 1. In Table 1, "–" means the related MCs do not only contain the corresponding POS, and different alphabet below "Class" means different classes.

By comparing the taxonomy of CCD and Cilin, we find that it is hard to achieve the corresponding relationship between CCD Category and Cilin Class/MC. For example, in Class level, Class "C", i.e., time and space, in Cilin covers two Categories in CCD such as "noun.Time" and "noun.Space", but MC "Aa" and "Ba" that are within different Cilin Classes like "person" and "thing" respectively. Another example is shown in Figure 1, we can see that the many-to-many mappings in MC level are much more complex than in Class level. However we find that all the concepts in the same subdivision in Cilin are almost corresponding to the same CCD category.

2.3 Organization

As same as the classification system, the organization of CCD follows WordNet as well, whose basic unit is CSynset. Words in a CSynset are considered as synonyms. CSynsets are divided into 45 categories and linked by semantic relations. All relations are tagged clearly and have various types. Some relations contain hierarchical information. If we consider CSynsets

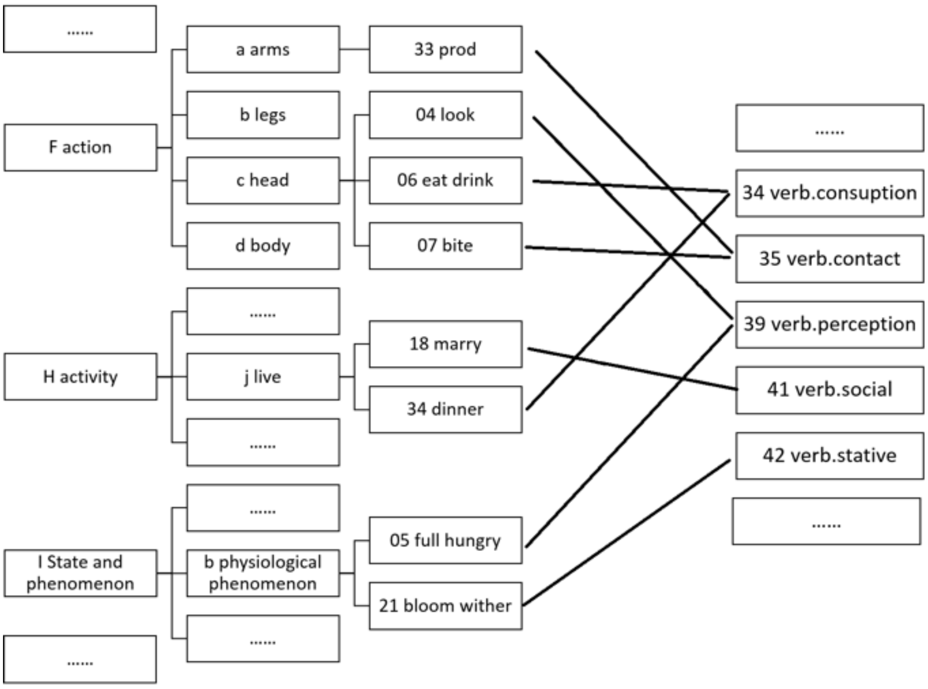


Figure 1: Example of relations between medium classes of Cilin and categories of CCD

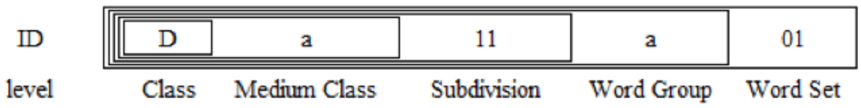


Figure 2: Category and ID format in Cilin

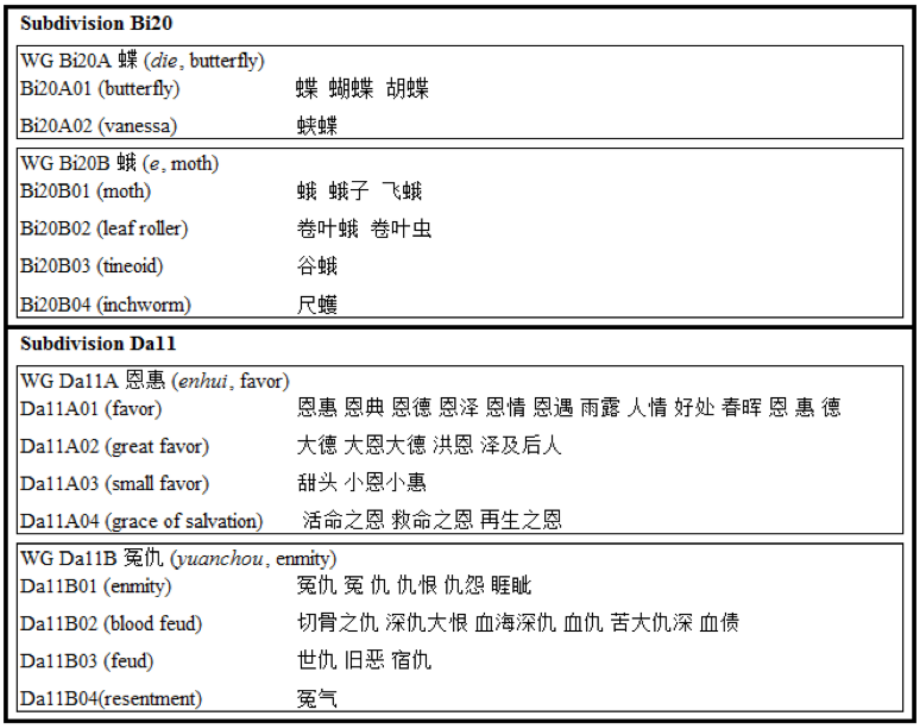
as nodes and hierarchical relations as edges, CCD can be represented as a forest-like graph.

The organization of Cilin is more close to the traditional thesaurus rather than CCD. A five-level hierarchical organization is preliminary defined by linguists, which are Class, MC, Subdivision, Word Group (WG) and Word Set (WS) from top to down. Figure 2 shows the ID format defined in Cilin, from which we can get the classification information of words. 80% of Cilin WSs are Synonym Sets, while the other 20% of them either consist of similar words or a single word, marked by "#" and "@" respectively. The first word in a WG is called title word (TW), which is the most representative one in the WS. The meanings of other words in the same WG are basically same as the TW, with some subtle differences in the emphasis of sense, emotional color or usage. These differences are represented by the division of word sets. So words in the same WS have a very high degree of semantic similarity.

Figure 3 shows some examples of WG. In Figure 3, the title word of each WG is noted after the WG id. For example, the title word of WG Bi20A is 蝶 (*die*, butterfly). The other WS in the Bi20A are either aliases (e.g., 蝴蝶 (*hudie*, butterfly)) or special cases (e.g., 蛱蝶 (*jiadie*, vanessa)) of *die*. In other WGs, the situations are similar.

In general, CCD shows clear and rich semantic relations but misses a large number of Chinese-specific words, while Cilin covers a wide range of Chinese words without clear relations between them. Therefore, we could build a better Chinese synonym resource by integrating the structure of CCD and the Chinese-specific words of Cilin together.

In order to achieve this goal, this paper proposes two methods. Firstly we ignore the inner structure of CCD and Cilin, and integrate them by direct mapping; Sec-



Subdivision Bi20	
WG Bi20A 蝶 (*die*, butterfly)	
Bi20A01 (butterfly)	蝶 蝴蝶 胡蝶
Bi20A02 (vanessa)	蛱蝶
WG Bi20B 蛾 (*e*, moth)	
Bi20B01 (moth)	蛾 蛾子 飞蛾
Bi20B02 (leaf roller)	卷叶蛾 卷叶虫
Bi20B03 (tineoid)	谷蛾
Bi20B04 (inchworm)	尺蠖
Subdivision Da11	
WG Da11A 恩惠 (*enhui*, favor)	
Da11A01 (favor)	恩惠 恩典 恩德 恩泽 恩情 恩遇 雨露 人情 好处 春晖 恩 惠 德
Da11A02 (great favor)	大德 大恩大德 洪恩 泽及后人
Da11A03 (small favor)	甜头 小恩小惠
Da11A04 (grace of salvation)	活命之恩 救命之恩 再生之恩
WG Da11B 冤仇 (*yuanchou*, enmity)	
Da11B01 (enmity)	冤仇 冤 仇 仇恨 仇怨 睚眦
Da11B02 (blood feud)	切骨之仇 深仇大恨 血海深仇 血仇 苦大仇深 血债
Da11B03 (feud)	世仇 旧恶 宿仇
Da11B04(resentment)	冤气

Figure 3: Examples of Word Groups in Cilin

ondly we consider the structure of CCD and Cilin, and integrate them by hierarchical mapping. The details of two methods will be discussed in next two sections.

3 Integration by Direct Mapping

We first present the direct mapping algorithm, which is a straightforward way of integrating heterogeneous thesauruses. The procedure of it is show in Algorithm 1.

Algorithm 1 Direct Mapping

Require: *Tong_dict* Cilin，*CCD_dict* CCD
Ensure: *merge_dict* the dict after merge

```
function MERGE_DICT(Tong_dict, CCD_dict)
Copy Tong_dict as Merge_dict
    for WS in Merge_dict do
        for w in WS do
            for CSynset in CCD_dict do
                if w in CSynset then
                    add all words in CSynset to WS of Merge_dict
                end if
            end for
        end for
    end for
    return Merge_dict
end function
```

Direct mapping takes Cilin and CCD as input (*Tong_dict* and *CCD_dict*, respectively), and go through all *WS* in Cilin and *CSynset* in CCD. It simply adds all the words in *CSynset* of CCD into *WS* of Cilin when two sets have the same word *w*. At last, the merged dictionary *Merge_dict* is returned by Algorithm 1 as output. The empirical performance of direct mapping is investigated in Section 5.

4 Integration by Hierarchical Mapping

CCD and Cilin have their own synonym sets, specifically, 99,646 CSynsets in CCD and 14,426 WS in Cilin. To integrate them, a mapping among these sets is needed. However, enumerating all possible mappings between tens of thousands of sets is quite inefficient. Instead, we propose to use the hierarchical structures of the two thesauruses to perform an efficient mapping.

Since both thesauruses are organized in a forest-based hierarchy, the mapping between higher structures could be used as the extra information to distinguish synonym sets. Therefore, we perform the hierarchical mapping in a top-down manner.

4.1 Assigning Cilin Subdivisions to CCD Categories

From Section 2, we can see that POS of words in CCD can be separated by their categories clearly, while in Cilin POS should be divided in subdivision level. So mapping classifications between CCD and Cilin should be done in subdivision level as well. In practice, we treat the subdivision as the largest group of semantic classes of Cilin, which will constrain the further search for synonyms. Because the lack of regularity, we map the 1,425 subdivisions onto CCD categories manually.

4.2 Mapping Word Groups and CSynset Subtrees

The mapping problem has been reduced to mapping WSs under a given subdivision to the CSynset in the sub-forest hierarchy of the corresponding category. This mapping is also difficult because a single word could appear in multiple synonym groups with different meanings. Instead of direct mapping, we propose to use larger unit WG and its local structure as a mapping constraint in order to perform disambiguation and improve efficiency. Our disambiguation relies on two important facts of Cilin:

- The general meaning of a WG is represented by its title word;
- WGs in a subdivision have similar meanings.

These facts could be briefly seen in Figure 3. All words in the same WG are either synonymy or hyponymy with its title word. And the title words of all WGs in the same subdivision either have the same hypernym or are antonyms.

We design an evaluation metric of the mapping between a WG and a CSynset subtree that has the following three criteria:

- **Criterion 1:** The root CSynset of subtree in CCD should contain the title word of the WG;
- **Criterion 2:** The CSynsets in subtree should cover the other words in the same WG as many as possible;
- **Criterion 3:** The semantic relations of CSynsets in CCD containing title words of WGs in the same subdivision should be as close as possible. The closeness of two CSynsets is measured by the number of semantic relations of hyponyms and antonyms on the path connecting them in CCD.

When a title word is not contained in any CSynset of CCD, we use the next word in the same WS to take its place. When no words in the WS are contained in CCD, we assign this WG as brothers of the mapped ones in the same subdivision. Most words in these WGs are language-specific concepts in Chinese.

Based on the above three constraints, we calculate the scores of possible mapping subtrees for each WG, and then choose the one which has the maximum score as our mapping result.

4.3 Mapping Word Sets and CSynsets

After the mapping of WGs and subtrees, the WS containing title word of a WG has already been mapped to the root CSynset of the corresponding subtree, and all the other WSs in the given WG should be mapped to CSynsets in the subtree. In this step, the words in a WS could be used to disambiguate different CSynsets.

Similar to the solution of mapping WGs and CSynset subtrees, we use the following two criteria to evaluate the mapping between WSs and CSynsets:

- **Criterion 4:** The CSynset should contain one or more words in the WS;
- **Criterion 5:** The CSynset should be close to the CSynsets that contains the other words in the same WS.

Based on the above two criteria, we calculate the scores of possible mapping CSynsets for each WS and choose the best mapping result with the maximum score. In particular, if the subtree of a WG has only one node, i.e., root, all WSs under the WG are added into the subtree as the direct child nodes of the root. If none of the words in a WS is contained by any CSynsets of the subtree, the WS is added into the subtree as a direct child node of the root as well.

We should note that the WSs marked by "#" in Cilin consists of similar words, which cannot put in the same CSynsets. As a result, we treat each word in these WSs as a new WS that contains a single word and map them onto CSynsets.

4.4 A Summary of the Proposed Method

In a nutshell, the key steps of the proposed hierarchical mapping approach to integrating heterogeneous thesauruses are summarized as follows:

1. Treating the subdivision as the largest group of semantic classes of Cilin, and assigning Cilin subdivisions to CCD categories;

下层社会 下层阶级
棍 杖 杆
声 语气 语调 声调 音调 调子 话音
计划 打算 方案 规划 筹划 盘算
雕版 梓
失陷 陷落
洋粉 洋菜 石花菜 营养琼脂 琼脂培养基
多动症 小动作癖
自由 无限制
游乐场 娱乐场 游艺场

Figure 4: An Example of Synonym Sets in new resource

2. Mapping word groups and CSynset subtrees according to Criterion 1-3, and selecting the mapping with the maximum score;
3. Mapping Word Sets and CSynsets according to Criterion 4-5, and selecting the mapping with the maximum score.

5 Empirical Study

5.1 Evaluation Method

After mapping, we get millions of synonym sets in our new resource, and it is difficult and not realistic to evaluate the result manually. Fortunately, we can evaluate the similarity between words by the distance of word vector. Word2Vec [10,11] maps a word to a dense vector with the fixed dimensionality. This method can retain some semantic information of words from their context. Word2Vec can also allow similar words in semantics be close to each other in vector space. For dataset, we use SogouCA[1)] to train the Chinese word embedding. Then, we use this word embedding to represent the word in synonym set generated from Section 3 and Section 4. We apply the average v_{avg} of all vectors in synonym set to represent the semantics of this synonym set. In other words, the average vector is the key word embedding in this synonym set. The Euclidean distance between this word and v_{avg} can be utilized to measure the possibility that a word belongs to this synonym set. The smaller the value, the higher the possibility. This calculation is show as Formula (1)

$$dis(v_i) = \|v_i - v_{avg}\|_2^2, \tag{1}$$

where v_i is a vector of word belonging the synonym set.

1) https://www.sogou.com/labs/resource/ca.php

Due to the scale of training dataset, we cannot cover all words in CCD or Cilin. For the $dis(\cdot)$ of this word, we use the average distance among the synonym set as their $dis(\cdot)$. Then, we sort the words according to $dis(\cdot)$ by ascending order, and get the nearest k as synonyms. The result is shown in Figure 4. Due to the different sizes of synonym sets, it is not convinced that using the same strategy to select the value of k. For both direct and hierarchical mapping, we use Formula (2) to get the value of k,

$$k = len(Synonym\_set) \times m + 2 \qquad m \in [0, 1], \tag{2}$$

where $len(Synonym\_set)$ denotes the length of synonym set, and multiplying m means we only select the top m. This strategy can get different k for different synonym sets. We will discuss the value of m in detail later.

5.2 Experimental Result

Table 2 and Table 3 represent the result of direct and hierarchical mapping. From Table 2, 125,146 synonym sets are got by hierarchical mapping, but we only get 66,815 sets by direct mapping. Direct mapping would cause the granularity of partitioned set cruder, since some same words may have different semantics. However, hierarchical mapping can retain this relation when considering the high-level mapping. Our method can keep the synonym sets which contain the same word in mind, and meanwhile has different semantics via using the high-level words.

There are 17,817 *WS* in Cilin and 99,642 CSynset in CCD, it seems that break the math formula of the union of two sets because our hierarchical mapping can get more than 117,459 synonym set. Actually, as mentioned in the last of Section 3, in hierarchical mapping, the *WS*s which begin as # in Cilin are not synonymous relationship, thus we get these words one by one to merge

Table 2: A Comparison of some metrics between two methods

	Direct	Hierarchical
# of synonym set	66,815	125,146
# of synonym set except single word	38,462	67,540
average # of words in synonym set	5.98	2.21
# of shared synonym set	58,332	
# of unique synonym set	8,517	66,814
average distance of unique synonym set	2.9868	2.1643

with CCD. This method will cause a lot of *single word*, which can be treated as a synonym set alone. From Table 2, hierarchical mapping can get more synonym sets compared with direct mapping except the single word.

We also count the average number of words in synonym set generated by two methods. It can be seen that the sets generated by direct mapping are larger since this method is straightforward and simple. We find that hierarchical mapping can cover 87.30% result of direct mapping from bottom 3 line of Table 2.

For the unique set, the average distance of unique synonym set shows that the hierarchical mapping can get closer distance in synonym set than direct mapping.

5.3 Effectiveness of different m

Table 3 shows the result when using different m in two methods. From Table 3, the average distance of hierarchical mapping is significantly less than the distance of direct mapping. It indicates that the result of hierarchical mapping is better than direct mapping. In other words, all words in synonym sets generated by hierarchical mapping is more likely to belong a same semantics. It is normal that the average distances of two method both increase when increasing m. However, average distance of hierarchical mapping among different m is more stable than that of direct mapping.

Table 3: Average distance for different m in two methods

m	direct mapping	hierarchical mapping
0.3	2.0710	1.7600
0.5	2.3101	1.8300
0.8	2.5837	1.9113
1.0	2.7585	1.9176

The above result indicates that the bad words in the hierarchical mapping have more confident than that in the direct mapping. This phenomenon could verify that the hierarchical mapping have higher stability than the direct one.

6 Conclusion

In this paper, we propose a hierarchical mapping method which can integrate two heterogeneous thesauruses, CCD and Cilin. Our method takes advantage of the organizing structures of two thesauruses and brings a novel integrated resource for Chinese synonyms. As a result, we obtain a novel resource with 125,146 synonym sets which covers 148,327 words. We use word2vec to evaluate our novel resource, and the experimental results show that the proposed hierarchical mapping can achieve better performance than direct mapping.

Acknowledgements. We would like to thank the anonymous reviewers for their insightful comments. Shujian Huang is the corresponding author. This work is supported by the National Science Foundation of China (No. U1836221, 61772261) and the Jiangsu Provincial Research Foundation for Basic Research (No. BK20170074).

References

1. C. Fellbaum, ed., *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press, 1998.
2. H. Liu and P. Singh, "ConceptNet: A practical commonsense reasoning toolkit," *BT Technology Journal*, vol. 22, no. 4, pp. 211–226, 2004.
3. H. Liu and P. Singh, "Commonsense reasoning in and over natural language," in *Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems*, (Wellington, New Zealand), pp. 293–306, 2004.

4. Z. Lu, "Expression of semantic relations and construction of knowledge system," *Language Application*, vol. 3, pp. 79–85, 1998.
5. J. Yu and S. Yu, "The structure of chinese concept dictionary," *Journal of Chinese Information Processing*, vol. 16, no. 4, pp. 12–20, 2002.
6. J. Mei, Y. Zhu, Y. Gao, and H. Yin, *TongYiCi CiLin*. Shanghai, China: Shanghai Dictionary Publishing House, 1996.
7. M. Carpuat, G. Ngai, P. Fung, and K. W. Church, "Creating a bilingual ontology: A corpus-based approach for aligning WordNet and HowNet," in *Proceedings of the 1st Global WordNet Conference*, (Mysore, India), 2002.
8. L. Mei, Q. Zhou, L. Zang, and Z. Chen, "Merge information in hownet and tongyici cilin," *Journal of Chinese Information Processing*, vol. 19, no. 1, pp. 64–71, 2005.
9. I. Niles and A. Pease, "Linking lexicons and ontologies: Mapping WordNet to the suggested upper merged ontology," in *Proceedings of the 2003 International Conference on Information and Knowledge Engineering*, (Las Vegas, Nevada), pp. 412–416, 2003.
10. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of the 1st International Conference on Learning Representations (Workshop Track)*, (Scottsdale, AZ), 2013.
11. N. Mrksic, D. Ó. Séaghdha, B. Thomson, M. Gasic, L. M. Rojas-Barahona, P. Su, D. Vandyke, T. Wen, and S. J. Young, "Counter-fitting word vectors to linguistic constraints," in *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (San Diego, CA), pp. 142–148, 2016.