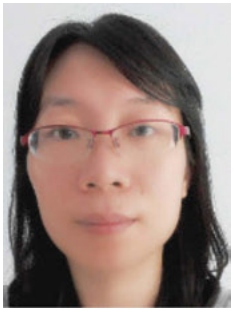
Improving Chinese Term Association from the Linguistic Perspective

Nan Li* and Jiqing Sun**

East China University of Science and Technology,
Room 308 Library, No.130 Meilong Road, Xuhui District, Shanghai 200237 P.R., China,
*<ajen@ecust.edu.cn>, **<jqsun@ecust.edu.cn>



Nan Li is a PhD candidate in management science and engineering at East China University of Science and Technology (ECUST). She holds an ME in computer science from ECUST and works at the Institute of Science and Technology Information of ECUST. Her research interests include knowledge organization systems, terminology, natural language processing, text mining and information retrieval.

Jiqing Sun is Professor at East China University of Science and Technology. Current research areas include knowledge management, information retrieval, technological innovation evaluation and patent analysis.

Li, Nan and Jiqing Sun. 2017. "Improving Chinese Term Association from the Linguistic Perspective." *Knowledge Organization* 44(1): 13-23. 43 references.

Abstract: The study aims to solve how to construct the semantic relations of specific domain terms by applying linguistic rules. The semantic structure analysis at the morpheme level was used for semantic measure, and a morpheme-based term association model was proposed by improving and combining the literal-based similarity algorithm and co-occurrence relatedness methods. This study provides a novel insight into the method of semantic analysis and calculation by morpheme parsing, and the proposed solution is feasible for the automatic association of compound terms. The results show that this approach could be used to construct appropriate term association and form a reasonable structural knowledge graph. However, due to linguistic differences, the viability and effectiveness of the use of our method in non-Chinese linguistic environments should be verified.

Received: 31 May 2016; Revised: 15 July 2016; Accepted: 15 August 2016

Keywords: semantic structure analysis, Chinese term association, morphemes, similarity, relatedness

1.0 Introduction

Terminologies have been developed to map the elements and rules of the world from a scientific perspective, and a large number of terms and their relations have been organized into a system to reveal domain knowledge. Therefore, the construction of a semantic association of terms is always the basic work for knowledge organization, information retrieval and bibliometrics. Until recently, the methods for automatic term association were still the focus and they mainly used various lexical resources or encyclopedias such as WordNet (Budanitsky and Hirst 2006), UMLS (Friedman et al. 2004) and Wikipedia (Gabrilovich and Markovitch 2007). However, these resources are mostly developed in English and mainly distributed in general areas or within several professional fields such as chemistry, biology and medicine. Unfortunately, there is a lack of such resources in non-English languages, and therefore it is difficult to achieve a sufficient number of knowledge bases for an application in a short time. Thus, the efficient identification of semantic relations without using external resources is an important topic.

From a linguistic perspective, several researchers have demonstrated the usefulness of exploiting the internal structure of words and modeling the various meaning-bearing units to improve semantic analysis. According to Saussure's structural linguistics theory, most compound words represent the form and semantic processing of their constituent morphemes. Therefore, morphological information is more frequently applied in semantic annotation, extraction and retrieval (Schulz and Hahn 2000; Zieman and Salas 2001; Mesfar 2010). These studies significantly focused on the word formation models and semantic features to achieve these applications, which make it possible

to explore the semantic relations of terms from a linguistic perspective.

It is no doubt that linguistic analysis is often closely connected with grammatical structures and linguistic phenomena; therefore, many studies have been conducted for various languages (Kupayeva 2015; Šojat and Srebačić 2014; Yang and Sun 2015). Chinese is a branch of the Sino-Tibetan language family in which the syntax and semantics differ from other languages and are mainly expressed by separate word formation and character sequence. Although it is reported that lexical semantics can be analyzed by morpheme structure and morpheme semantic combination (Lu 1957; Qiu 2006), we believe that there is still room to find its potential applications in a specific domain. For example, the semantic role and function of morphemes in different domain terms are significantly different, such as "聚" (poly-), "乙" (di-) and "烯" (alkene) in "聚乙烯" (polythene). Instead of simply decomposing the morpheme structure, we intended to model the semantic relations to better distinguish the semantic types of morphemes.

In this work, the existing method for automatic semantic association was improved by using morpheme parsing. Based on the consistency between Chinese word formation and its meaning from a linguistic perspective, we have designed a morpheme-based term association model for a Chinese knowledge organization system. The structure of this article is organized as follows: The next section reviews the related work in semantic association and Chinese lexical semantic computation. Then we present a term association model based on Chinese morphemes. Finally, the procedure and results of an experiment are described, followed by the conclusion.

2.0 Review of related literature

2.1 Semantic association

Semantic association refers to the construction of complex relationships between concepts or entities, the basic units in knowledge organization (Hjørland 2003). There are two different relations: the similar relation, e.g., the hierarchical and equivalence relations, and the correlative relation (Bräscher 2014). The similar relation reflects the continuity of knowledge based on the similarity of their meaning, and the correlative relation emphasizes the nonsimilar logical connection. These relations can be estimated using semantic similarity or semantic relatedness.

2.1.1 Semantic similarity

Many works have been written over the last few years proposing different ways to measure semantic similarity. Among them, the lexical-based measure is a typical ontology-independent approach. The lexical similarity measures began with the heuristic homology algorithm of Smith and Waterman (1981). They first introduced a method for calculating the maximum similar element of a textual sequence. Bourigault (1999) proposed a term extraction tool, LEXTER, decomposing the multi-word term into two syntactic constituents (head and expansion), and the approach was widely used (Assadi 1997; Drymonas et al. 2010; Zhang et al. 2009) to build the semantic relations of concepts, particularly similarity measure. Similar studies have been conducted in the scientific field. For example, Klinger et al. (2008) developed a method for identifying IUPAC (a nomenclature for organic chemistry recommended by the International Union of Pure and Applied Chemistry) and IUPAC-like chemical names by finding the structural classes, atoms and elements, which are the fragments in IUPAC representations. The Chinese morpheme-based method belongs to this type of measure such as the single-character-based similarity algorithm (Zhu et al. 2002) described in the following section. The aforementioned works mostly used a formalized rule to obtain the semantic element and then applied it to identify or extract terms and their relations. Despite the rule restrictions, lexical-based measures are still highly feasible and effective to automatically identify the semantic relation without a knowledge base or corpus. However, the existing research is better suited for the explicit grammar features of Indo-European languages. In contrast, Sino-Tibetan languages are more complex for linguistic rules. More applied research should be conducted for different languages, particularly for Chinese.

2.1.2 Semantic relatedness:

Semantic relatedness is a metric method using statistical means to correlate terms such as path-based measure (Hirst and St-Onge 1995), gloss-overlapping measure (Banerjee and Pedersen 2003), and co-occurrence measure (Patwardhan and Pedersen 2006). Whatever the approach may be, the topic of effective relatedness in semantics is important, as it shows how to decide whether those co-occurring terms do in fact have close ties or whether they just appear together. Many improved methods for semantic relatedness focus on distinguishing the knowledge connection between concepts. Most of these approaches attempt to directly determine a strong or weak connection by its statistical strength. Zhang et al. (2012) selected the co-occurrence frequency of keyword pairs to filter the less common relatedness. Hu and Chen (2014) assumed that the reliable connection not only appears more frequently, but also occurs in various documents. Therefore, they used the combined word and document frequency as a connec-

tion strength to filter weak connections. However, a different opinion stated that the word or document frequency method is not enough, and the characteristic features of the context should also be included. Kwon (2014) proved that the betweenness centrality, term frequency, effective size and complexity of a subject affect the number of a semantic relation. Wu and Zhao (2008) used the number of times cited as the attribute of the article to implement a weighted co-word model. Another method for semantic co-word analysis was attempted by extracting keywords from full texts (Wang and Wang 2014). Furthermore, the lexical, syntactic and semantic features such as the syntax template or the contextual graph could also be used to build semantic association (Hearst 1992; Bounhas 2011; Girju et al. 2014). All these studies show that a contextual model contributes to the effectiveness of semantic relatedness, and the effectiveness of context feature is one of the most important challenges. Although the word-level semantic analysis has many weaknesses, we introduce a new contextual model at the morpheme level.

2.2 Chinese lexical semantic computation

The Chinese character is the basic unit of Chinese in grammar, whereas the Chinese morpheme referring to an entire independent meaning corresponds to one or more characters. Several linguistic studies on the lexical morpheme structure or meaning of morpheme have been carried out (Lu 1957; Qiu 2006); they can be used to explore semantic association from the linguistics perspective. In 1999, the formation of Chinese characteristics was first used to evaluate word similarity, known as "literal-based similarity algorithm." Based on this, Hou and Wu (2001) tracked the performance of a word-element-based similarity algorithm (also called morpheme-based similarity) and a single-character-based similarity algorithm in which Chinese word structure and Chinese expression rule were also considered. For example, it has been stated (Zhu et al. 2002; Hou and Wu 2001) that "In Chinese, the core meaning is always located at the end of conceptual representation" which is the semantic core principle. Many studies (Zhang 2005; Ran and Sun 2011; Chang and Zhang 2012) improved the literal-based similarity algorithm with a combination of semantic lexicon and statistical features. Only the count and frequency of word units were used in the aforementioned methods, but most of them ignore the language function and semantic characteristics of morphemes. Therefore, the "literal" analysis of Chinese strings could not map the meaning of the terms; this was the major limitation of similar studies.

In this study, we extend our earlier work by first proposing a new and effective way of Chinese term association using morpheme-based semantic analysis in which specific-domain morphemes were collected and classified by their functions. Then, this method was combined with a single-character-based similarity algorithm and a reliable relatedness algorithm to improve the overall performance.

3.0 Morpheme-based term association model

To connect the specific-domain terms in a semantic and automatic way, this study proposes a morpheme-based term association model and introduces an integrated method for improving the semantic similarity and relatedness algorithm by semantic structure analysis at the morpheme level.

3.1 Model definition

Each normalized specific-domain term can be viewed as a structured morpheme sequence with a special formation pattern. Therefore, a semantic analysis of terms is based on the cognition of specific morphemes including their stability, specialization and diversity (Li et al. 2015). According to this concept, our term association model is defined as follows.

3.1.1 Definition

Given a collection of morphemes C, there is a type label t_i for each single morpheme c_i denoting the function type of a morpheme according to the significance of the concept. Given a set of specific-domain terms W, the term w_i can be viewed as a sequence of morphemes q_i, i.e., $[(w_i \rightarrow q_i = \{(c)]_1\ c_2 \dots c_j \dots) \mid c_j \in C\}$. The characteristics of the sequence differ from the types, numbers and positions of constituent morphemes. Then, the semantic association of terms was obtained by estimating the similarity of the above mentioned features, namely, $R(w_i,w_j) = f(q_i,q_j)$. Moreover, term association is a weighted combination of semantic similarity (R_{sim}) and semantic relatedness (R_{rel}); this can be expressed as Equation (1):

$$\mathrm{R}(W_i, W_j) = \{R_{sim}(W_i, W_j) \cup R_{rel}(W_i, W_j)\} = \alpha \cdot f_{sim}(Q_i, Q_j) + \beta \cdot f_{rel}(Q_i, Q_j)$$

3.1.2 Phases

The process of term association can be divided into three phases, as shown in Figure 1. The first step is "morpheme parsing" (S1), aimed at establishing morpheme sequence mapping (q_i) for each term (w_i) in sets W and extracting the formation mode of term meaning. This step significantly prepares the available semantic units or morphemes for the next step. In the "semantic computa-

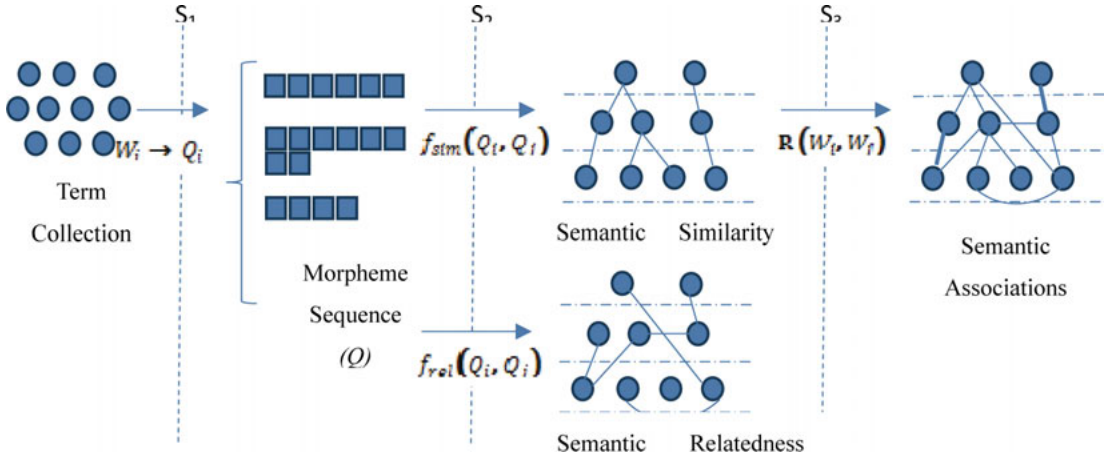


Figure 1. Three phases of the term association process.

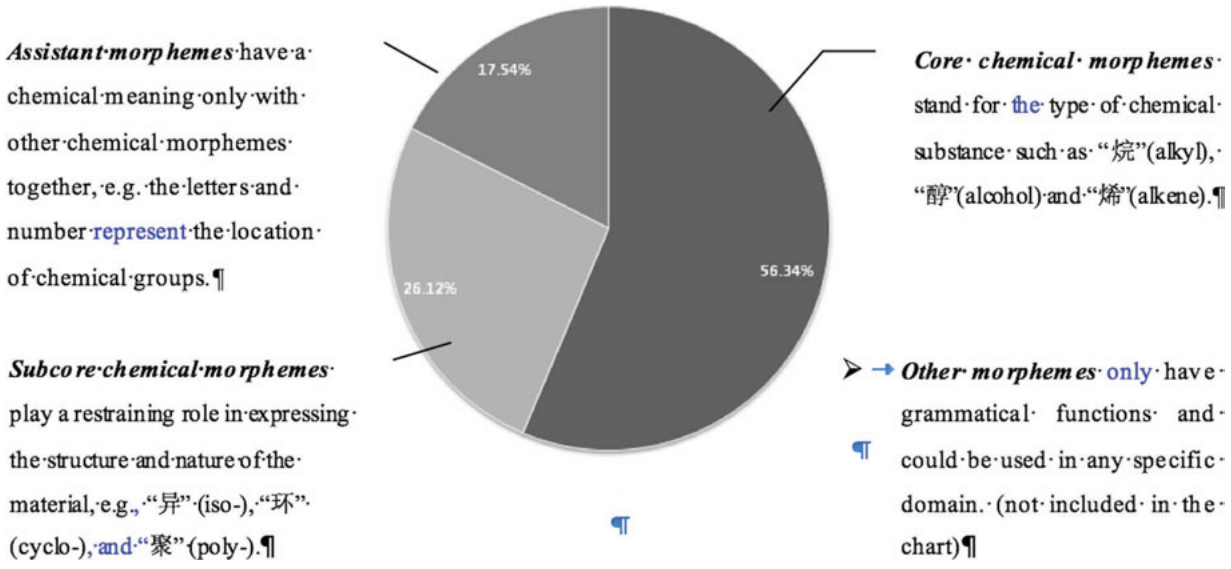


Figure 2. Composition of chemical morpheme collection

tion" step (S2), the semantic similarity and relatedness were separately calculated using the morpheme sequence. The final step (S3) combines the aforementioned two relations with the entire semantic association structure of terms.

3.2 Methods

3.2.1 Morpheme parsing

Morpheme parsing is based on a systematic and comprehensive morpheme set with a slight change and is only suitable for specific areas (Li et al. 2015). A morpheme set is built in two steps: 1) by verifying candidate morphemes; and 2) by classifying them. For example, we obtained 268 Chinese chemical morphemes based on chemical name specification and expertise and divided them into four groups, namely, core morphemes (A), subcore morphemes (B), assistant morphemes (C) and others (D), according to the importance of morpheme in concept expression and language function (see Figure 2).

According to morpheme collection, chemical terms can be translated into morpheme sequence or morpheme-type-label sequence, e.g., "聚乙烯"(polythene) can be expressed as "{聚(poly-)|乙(di-)|烯(alkene)}" or "{B|C|A}" in Chinese morpheme formation (Li et al. 2010). Following these methods, the semantic structure of each term can be parsed into such a formation and quantify the semantic content of terms by rules of formation, such as the knowledge value of the term ("K-value"). Based on Shannon's information theory, the entropy of an information source can be calculated using the probability mass function of each source symbol to be communicated. Then, if the morpheme formation of the term is considered, the

"K-value" of terms is given by the sum of the weighted entropy of various morphemes, defined as follows:

$$k\_value(w_i) = -\sum_{t_i} \alpha_i * p(t_i) * log_2\, p(t_i)$$
$$i = 1,\ 2,\ \ldots\ldots$$

Here, $p(t_i)$ is the probability of occurrence of the **i**-th possible value of morpheme type t_i, and each t_i has a corresponding weighting coefficient α_i according to the importance of this morpheme type in term. The k-value quantifies the contents of domain-specific term and reflects its specific degree or location in the knowledge hierarchy where the term in the higher level is more abstract and has a smaller k-value in general.

3.2.2 Semantic similarity measure

So far, the literal-based similarity algorithm is still a popular algorithm, which matches the basic lexical unit such as characters, morphemes and concepts to judge the similarity of words. Considering the semantic function of morphemes, different features were added to the semantic similarity measure. Several assumptions were made as follows:

- If two terms have more of the same morphemes, they are more similar.
- If there are more core morphemes in a set of matched morphemes, the two terms are more similar.
- According to the semantic core principle, the matched morpheme located in a rear position plays a more important role than others.
- The difference in term length can reduce the similar probability of terms. Hence, the ratio of term length was introduced as a parameter.

In similarity metrics, three factors were considered, 1) the common matched morphemes, 2) term length, and 3) morpheme position, to improve the literal-based similarity algorithm, defined as follows:

$$R_{sim}(w_i, w_j) = \alpha * match(w_i, w_j) + \beta * len_{ratio(w_i,w_j)} * pos(w_i, w_j)$$
$$\alpha + \beta = 1; \alpha, \beta > 0$$

Here $match(w_i,w_j)$ calculates the average proportion of the common k-type morpheme c_k in two terms $\mathbf{w_i}$ and $\mathbf{w_j}$. $Len_{ratio(w_i,w_j)}$ uses the ratio of term length as the term-length coefficient, which should be <1, and $pos(w_i,w_j)$ computes the position weight of each common k-type morpheme. In equation (3), weight α and β were assigned empirical values of 0.6 and 0.4, respectively.

Equations (4)-(6) are as follows:

$$match(w_i, w_j) = \sum_k p_k * \frac{\left[\frac{count(c_k)}{len(w_i)} + \frac{count(c_k)}{len(w_j)}\right]}{2}$$

$$len\_ratio(w_i, w_j) = \begin{cases} \frac{len(w_i)}{len(w_j)} & if\ (len(w_i) < len(w_j)) \\ \frac{len(w_j)}{len(w_i)} & if\ (len(w_i) > len(w_j)) \end{cases}$$

$$pos(w_i, w_j) = \left(\sum_k p_k * \left[\frac{locate(c_k)}{\sum locate(w_i)} + \frac{locate(c_k)}{\sum locate(w_j)}\right]/2\right)$$

Here, p_k is assigned to reflect the core level of k-type morpheme c_k. $Count(c_k)$ and $len(w_i)$ separately measure the number of all k-type morphemes and the length of term w_i, whereas $locate(c_k)$ or $locate(w_k)$ stands for the location of each k-type morpheme.

3.2.3 Semantic relatedness measure

To ensure the validity of term relatedness, a novel context-based approach using two morpheme descriptors as the contextual features was introduced. In the remainder of this section, the co-occurrence relatedness method improved by our approach is described.

First, the co-occurrences of two terms and the occurrences of a single one in a sentence were separately counted, and the strength of association $R_{rel}(w_i,w_j)$ between the two terms was measured according to Jaccard coefficient, defined as follows:

$$R_{rel}(w_i, w_j) = \frac{sc(w_i, w_j)}{sc(w_i) + sc(w_j) - sc(w_i, w_j)}$$

where sc is the number of sentences where one or two terms appear.

Then, no more limitation was observed for co-occurrence association, except for the range appearing in the sentence. The effectiveness of relations can be measured with context features; therefore, this paper proposes a specialization level and context similarity to filter the unreliable co-occurrence associations.

3.2.3.1 Specialization level

Generally, the more frequent is the appearance of two terms in scientific literature, the higher is the probability of semantic correlation from a professional perspective. Moreover, the context of the literature presents the technical terminology, whose analysis can be conducted at the morpheme level. Therefore, we propose an indicator $Context_{spec}$ to measure the specific morpheme content of the context, indicating the specialization level of scientific literature as follows:

$$Context\_spec(w_i, w_j) = avg(count(c_k)/len(s_{ij}^k))$$

Here, s_{ij}^k stands for the sentence k where term w_i and w_j appear together, and c is the specific morpheme used in sentence k. $Count(c_k)$, $len(w_i)$, separately measure the number of all specific morphemes in sentence k and the length of sentence k. The specialization level index $Context_{spec}$ is the average share of the component morpheme quantity in a sentence.

3.2.3.2 Context similarity

The correlation between terms varies directly as the co-occurrence frequency of terms, i.e., if the same term pair appears in different articles, they are more likely related to each other. Notably, the variety of co-occurrence context information is important for relationship analysis because the co-occurrence in different contexts means a higher chance of semantic association than in a similar situation. A specific morpheme was still selected as a semantic feature described in the context, and the morpheme sequences of context were compared. The context similarity $Context_{sim}$ between terms w_i and w_j is the average similarity of any two morpheme sequences of context described as follows:

$$Context\_sim(w_i, w_j) = avg(\forall sim(s_{ij}, s_{ij}'))$$

$$sim(s, s') = k/n$$

where s_{ij} or s_{ij} is the morpheme sequence of context where terms w_i and w_j co-occur, and the sequence similarity $sim(s,s')$is the ratio of the length of the same morpheme sequence k to that of max morpheme sequence n (Smith and Waterman, 1981).

The aforementioned two indicators can be normalized and used to adjust the association strength $R_{rel(\mathbf{wi},\mathbf{wj})}$. The equation is as follows:

$$R_{rel}'(w_i, w_j) = R_{rel}(w_i, w_j) * Context\_spec'(w_i, w_j)/Context\_sim'(w_i, w_j)$$

4.0 Experiments

This section describes the term-association experimental examples of Chinese chemical substance terms and is organized as follows: the section on "experiment" lists the source of data and the algorithm used for the experiment; the section on "results of the experiment" shows the chemical-term associations using graphs and discusses the accuracy and effectiveness of the experimental results.

4.1 Experiment

Chemical substance terms were selected from the Chinese Science Citation Database (CSCD) for use as the test collections to validate our term-association approach. The experiment included the random selection of chemical articles and valid chemical-term filter from the keywords. To compare the effectiveness of our method in different datasets, 200 articles were selected as the control group (D_2) from a basic group of 400 articles (D_1), i.e., $D_2 \in D_1$. Within the two groups, there are 834 and 509 valid terms, respectively, which were filtered by the specific morpheme structure ratio of keywords.

Based on section 3.2, the algorithm used in the experiment was as follows:

- Step 1: Measure the knowledge value k_{value} of each term w_i in set W. The term of the higher knowledge value is the upper concept in the term system, and the smaller knowledge value is the lower concept.
- Step 2: Add virtual nodes for the term system to ensure the correctness of term association. There is usually a lack of appropriate linkable terms in a small term set, for example, terms "乙二醇" (ethylene glycol) and "乙醇" (ethyl alcohol) are similar in Chinese word form, but belong to diols and monohydric alcohols, respectively. Hence, "醇" (alcohols) and "二醇" (glycols) were added into term system. "二醇" (glycols) was connected with the narrower term "乙二醇" (ethylene glycol), and "醇" (alcohols) was set as the boarder term of "二醇" (glycols) and "乙醇" (ethyl alcohol). In particular, the added terms can be automatically extracted from existing terms by identifying the sequence of specific morphemes in the term. Subsequently for each added term, repeat Step1 until all the terms have their k_{value} s.
- Step 3: Calculate the similarity R_{sim} of each term w_i with the other term w_j, which should satisfy the condition that $k_{value(wi)} > k_{value(wj)}$, i.e., term w_j is narrower than w_i. Based on the semantic core principle, if the sequence of specific morphemes at the end of term w_i is the same as that of term w_j, a direct connection would be built with priority. This is called semantic-core matching. During the processing, when there is a similarity between more than one term $\mathbf{w_j}$, term $\mathbf{w_j}$ with the maximum similarity should be connected. Repeat this process until every node has at least one connection.
- Step 4: Calculate the co-occurrence relatedness R_{rel} between term pairs and optimize R_{rel} with specialization level and context similarity. The result only maintains a part of the connections, whose association value is larger than the experiential threshold ε. We considered ε = 0.015 because preliminary experiments show that for ε

> 0.015, the relatedness can be accepted by domain experts.

– Step 5: Combine two types of connections to the final result $R = R_{sim} \cup R_{rel}$, and the degree of association is defined as follows:

$$R = \alpha \cdot R_{sim} + \beta \cdot R_{rel}, \quad \alpha = 0.6, \ \beta = 0.4$$

4.2 Results of the experiment

This section describes the results of the experiment with examples of chemical-term association, as shown in Table 1.

Number	$D_{1(400)}$	$D_{2(200)}$
Keywords	1447	847
Initial chemical terms	834	509
Supplementary chemical terms	163	120
Connections based on semantic similarity	950	594
Connections based on semantic relatedness	1238	829

Table 1. Summary of experiment results.

Using the Gephi software, the experiment results were visualized, and two knowledge graphs of chemical terms in D_1 and D_2 were produced. As shown in Figure 3, the two term networks express the relationships of Chinese chemical term nodes used in this study. There were made up of the initial terms selected from the keywords in D_1 and D_2 and supplementary terms as needed. The term networks are complex, even though only 400 articles were considered, and the two graphs had multi-center structures and similar topological properties when the conventional social network analysis method was used. Thus, in this study, the term networks were filtered to show all the core term clusters consisting of a single core node and its associated nodes.

Figure 4 shows the details of part alcohol term cluster. Term "醇" (alcohols) is the core node in this view, and there are several subcore terms such as "甲醇" (methyl alcohol) and "乙醇" (ethyl alcohol), directly connecting to the core node. Moreover, the narrower terms of "二醇" (glycols) include"乙二醇" (ethylene glycol), "丙二醇" (propylene glycol) and "丁二醇" (butylene glycol). Each direct connection between two terms is built with similarity and relatedness measures, and there is a greater distance of each term node from the local root with an increase of "k-value." Besides, Figure 4 also shows some non-alcohol terms, for example, alkanes, which are mainly linked by co-occurrence relatedness. In conclusion, the knowledge graph has a reasonable structural layout according to general domain knowledge, and the experiment results demonstrate the effectiveness of our approach.

4.2.1 Accuracy of the morpheme-based term association model

The proposed model was validated by collecting expert opinions as to whether the connection between terms is correct, including: 1) accuracy of semantic similarity (P_1) for evaluating the rationality of knowledge hierarchical structure and the correlation between similarity metrics and semantic relation intensity in a specific domain; 2) accuracy of semantic relatedness (P_2) for estimating chemical relationships divided into two groups, related, i.e. reaction-related, property-related, and nonrelated, i.e. two terms co-occur in the same context, but have no direct links from a chemical perspective; and, 3) accuracy of all the relationships between terms (P) for all the types of relations.

Table 2 shows that the performance of our approach is at an above average level with values of 76.67% for D_1 and 76.74% for D_2. The results also show that the correct rates for D_1 and D_2 based on semantic similarity were 85.03% and 81.37%, respectively, whereas the correct rates for D_1 and D_2 based on semantic relatedness were 73.23% and 73.30%, respectively. The accuracy of semantic similarity was 10% higher than that of semantic relatedness, indicating that the similarity measure based on morpheme parsing is more suitable for the construction of a hierarchical knowledge system.

	P_1	P_2	P
D_1	85.03%	73.23%	76.67%
D_2	81.37%	73.30%	76.74%

Table 2. Experiment performance on Chinese chemical terms.

With regard to the semantic similarity measure, our approach follows the rule of chemical term formation and is consistent with domain knowledge. However, the performance of our method depends on two factors:

1) The effective combination of the formative rules for similarity algorithm. Because of the fact that the basic literal-based similarity algorithm easily leads to an incorrect similarity value, e.g., "苯氧基乙醇"(phenoxyethyl alcohol) was connected with "苯乙醇"(phenethyl alcohol) by a high similarity value, and the matching algorithm of the semantic core was designed to correct the mistake by linking "苯氧基乙醇"(phenoxyethyl alcohol) and "苯乙醇"(phenethyl alcohol) to "乙醇"(ethyl alcohol) separately. The matching algorithm of the semantic core established the direct connection by matching the sequence of specific morphemes at the end of the term. The single sequence of term ending position was considered in the above mentioned experiment. In fact, many lengthy chemical names contain a multi-

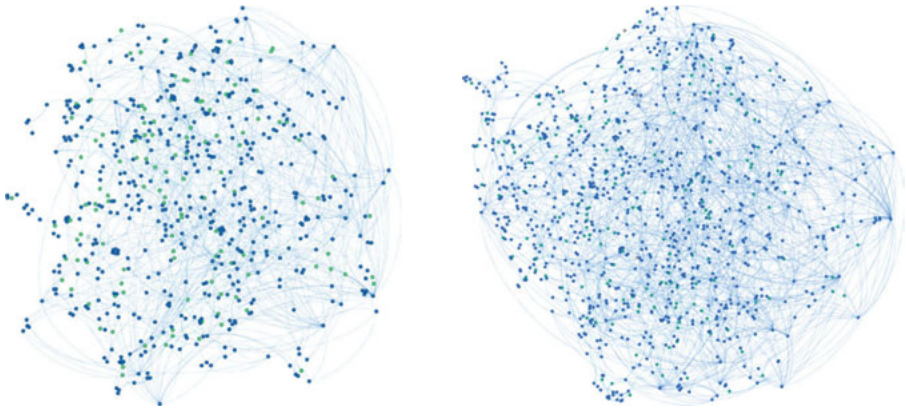


Figure 3. Knowledge graph of Chinese chemical substance names

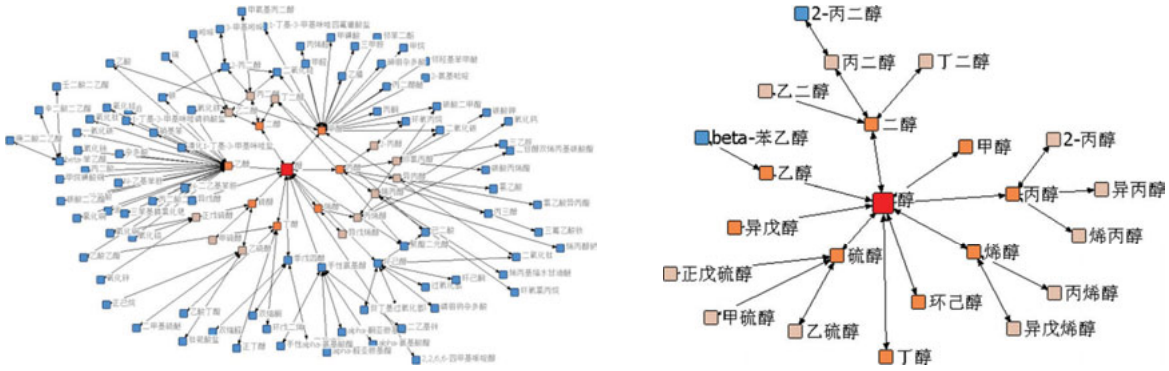


Figure 4. Detail view on knowledge graph of Chinese alcohols terms.

nesting structure; therefore, the multicore formation of a chemical term should be of concern.

2) The construction of a comprehensive set of specific morphemes. The semantic structure analysis at the morpheme level is the basis of our approach. As a fundamental knowledge resource, this is essential to predefine the specific morpheme sets for special subjects. For a subject like chemistry, there are numerous compound chemical names including substance, reactions and properties. The different sets of a specific morpheme should be built for each type of chemical name, as our experiment used a user-built chemical morpheme set as described in section 3.2.1. Specifically, the type of morpheme has an important effect on semantic metrics. For example, if "炔"(alkyne) is classified as a part of the core chemical morpheme, the "K-value" of "炔醇" (alkynol) is 0.8844; otherwise, the value decreases to 0.6744 when "炔"(alkyne) is viewed as a subcore chemical morpheme. Obviously, the similarity metrics will be affected by the variation in the "K-value," and the connection between terms will also be different. Therefore, whether the morpheme collection is complete and has a rational classification must always be the most important thing for consideration.

4.2.2 Effectiveness of term association

According to the results of the experiment (as seen in Table 3), the overlap ratio of semantic associations in two datasets was as high as 93.46%. Almost all the semantic relatedness links in D_2 appeared in D_1, and > 80% of the semantic similarity links in D_2 also appeared in D_1. The main reason for these results is that the term nodes in D_1, but not in D_2, affected the similarity metrics and changed the direction of term association, whereas the relatedness measure based on co-occurrence only depends on whether the node exists. Despite this, the majority of semantic links have not changed and the entire structure of term association is stable just as illustrated in Figure 3. All the results show that the data have less impact on the experiment results and our approach is relatively stable for term association.

As shown in Figure 3, two different colors represent the original term node (blue) and virtual term node (green). By comparing the integrated colors of the two graphs shown in Figure 4, a significant trend was observed: The virtual term node numbers decreased, even though the scale of the original term nodes increased. Moreover, the percentage of virtual term nodes decreased from 23.57% in D_2 to

	Total connections	Connections based on semantic similarity	Connections based on semantic relatedness
Occur in D_2	1423	594	829
Co-occur in D_1 and D_2	1330	503	827
Overlap Ratio	93.46%	84.68%	99.76%

Table 3. Overlap ratio of semantic associations.

19.54% in D_1, and the proportion of connection through virtual terms shrank by 1%. Thus, it can be concluded that the semantic association becomes more complete with increasing scale of original term nodes, and in practice, when the scale of original terms is limited, the virtual term should be added to the node set to ensure the accuracy of the connections.

5.0 Conclusions

The goal of this project is to promote a new solution for knowledge association by semantic structure analysis at the morpheme level. Using the literal-based similarity algorithm and co-occurrence relatedness method, this article reports a Chinese morpheme-based term association model and validates its performance by an experiment. The results indicate that it is very helpful to utilize the language function and the semantic role of Chinese morpheme, particularly by applying semantic structure analysis to enhance the efficiency of the semantic computation. This makes our approach feasible for the automatic association of compound terms. There is no doubt that multiple algorithm fusion makes term association more precise and comprehensive. The context and the case that this article presents should certainly contribute to the improvement of knowledge organization methods. We hope that this study will provide a better solution for automatic knowledge organization by combining and improving various algorithms.

References

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Cherry, J., Davis Michael, P. Allan, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin and Gavin Sherlock. 2000. "Gene Ontology: Tool for the Unification of Biology." *Nature Genetics* 25, no. 1: 25-9.

Assadi, Houssem. 1997. "Knowledge Acquisition from Texts: Using an Automatic Clustering Method Based on Noun-Modifier Relationship." In *Association for Computational Linguistics: Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, July 7-11 1997, Madrid, Spain,* ed. R. Mitkov and B. Boguraev. 504-6.

Banerjee, Satanjeev and Ted. Pedersen. 2003. "Extended Gloss Overlaps as a Measure of Semantic Relatedness." In *American Association for Artificial Intelligence: Proceedings of the 18th international joint conference on Artificial intelligence August 9-15 2003, Acapulco, Mexico,* ed. Anthony G. Cohn. 805-10.

Bodenreider, Olivier. 2004. "The Unified Medical Language System (UMLS): Integrating Biomedical Terminology." *Nucleic Acids Research* 32: 267-70.

Bounhas, Ibrahim, Bilel Elayeb, Fabrice Evrard and Yahya Slimani. 2011. "Organizing Contextual Knowledge for Arabic Text Disambiguation and Terminology Extraction." *Knowledge Organization* 38: 473-90.

Bourigault, Didier and Christian Jacquemin. 1999. "Term Extraction+ Term Clustering: an Integrated Platform for Computer-Aided Terminology." In *Association for Computational Linguistics: Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics, June 8-12, 1999, Bergen, Norway,* ed. H. Thompson and A. Lascarides, 15-22.

Bräscher Marisa. 2014. "Semantic Relations in Knowledge Organization Systems." *Knowledge Organization* 41: 175-80.

Budanitsky, Alexander and Graeme Hirst. 2006. "Evaluating WordNet-Based Measures of Lexical Semantic Relatedness." *Computational Linguistics* 32, no. 1: 13-48.

Chang, Xiaolong and Hui Zhang. 2012. "Construction of Chinese Polarity Lexicon by Integration of Morpheme Features." *Journal of Computer Applications* 32, no. 7: 2033-7.

Dong, Zhendong and Qiang Dong. 2001. "Construction of a Knowledge System and its Impact on Chinese Research." *Contemporary Linguistics* 3, no. 1: 33-44.

Drymonas, Euthymios, Kalliopi Zervanou and Euripides G. M. Petrakis. 2010. "Exploiting Multi-Word Similarity for Retrieval in Medical Document Collections: The TSRM Approach." *Journal of Digital Information Management* 8, no. 5: 315-21.

Friedman, Carol, Lyudmila Shagina, Yves Lussier and George Hripcsak. 2004. "Automated Encoding of Clinical Documents Based on Natural Language Processing." *Journal of the American Medical Informatics Association (JAMIA)* 11, no. 5: 392-402.

Gabrilovich, Evgeniy and Shaul Markovitch. 2007. "Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis." In *Proceedings of the*

20th International Joint Conference on Artificial Intelligence, January 9-12 2007, Hyderabad, India, ed. Ramon Lopez de Mantaras*,* 6: 1606-11.

Girju, Roxana, Brandon Beamer, Alla Rozovskaya, Andrew Fister and Suma Bhat. 2010. "A Knowledge-Rich Approach to Identifying Semantic Relations between Nominal." *Information Processing and Management* 46: 589-610.

Hearst, Marti A. 1992. "Automatic Acquisition of Hyponyms from Large Text Corpora." In *Association for Computational Linguistics: Proceedings of the 14th International Conference on Computational linguistics (COLING '92), August 23-28 1992, Nantes, France,* 2: 539-45.

Hirst Graeme, St-Onge David. 1995. "Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms." *Lecture Notes in Physics* 728, no. 9: 123-49.

Hjorland, Biger. 2003. "Fundamentals of Knowledge Organization." *Knowledge Organization* 30: 87-111.

Hou, Hanqing and Zhiqiang. Wu. 2001. "Using Single-Character-Based Similarity to Identity Chinese Cognate Words." In *China Society for Scientific and Technical Information: Proceedings of the 15th conference on computer information management, August 6-9 2001, Hai la'er, China*, 222-9.

Hu, Changping and Guo Chen. 2014. "An Exploration of Hierarchical Domain Knowledge Network and Its Micro-morphology Based on Co-word Analysis with Reliable Relations." *Journal of the China Society for Scientific and Technical Information* 33, no. 2: 130-9.

Klinger, Roman, Corinna Kolářik, Juliane Fluck, Martin Hofmann-Apitius and Christoph M. Friedrich. 2008. "Detection of IUPAC and IUPAC-like Chemical Names." *Bioinformatics* 24, no. 13: 268-76.

Krötzsch, Markus, Denny Vrandecic, Max Völkel, Heiko Haller and Rudi Studer. 2007. "Semantic Wikipedia." *Web Semantics Science Services and Agents on the World Wide Web* 5, no. 4: 251-61.

Kupayeva, A. K. 2015. "Word Formation Models and Semantic Features of Derived Words in Orhon Inscriptions (Derivations of Nouns and Adjectives)." *TRAMES Journal of the Humanities and Social Sciences* 19, no. 2: 171-88.

Kwon, Sun-young. 2014. "A Study on the Factors Influencing Semantic Relation in Building a Structured Glossary." *Journal of the Korean Society for Library and Information Science* 48, no. 2: 353-78.

Li, Nan, Jiqing Sun, Jiuming Ji and Rong Chen. 2015. "Knowledge Association Method Based on Semantic Relatedness." *Journal of the China Society for Scientific and Technical Information* 34, no. 6: 608-15.

Li, Nan, Rongting Zheng, Jiuming Ji and Qingqing Teng. 2010. "Research on Chinese Chemical Name Recognition Based on Heuristic Rules." *New Technology of Library and Information Service* 5: 13-7.

Lu, Zhiwei. 1957. *Study on Chinese Word Formation.* Beijing: Science Press.

Mesfar, Slim. 2010. "Towards a Cascade of Morpho-Syntactic Tools for Arabic Natural Language Processing." In *Proceedings of the 11th international conference on Computational Linguistics and Intelligent Text Processing, CICLing 2010, March 21-27, 2010, Iasi, Romania,* 150-62.

Nenadić, Goran, Irena Spasić and Sophia Ananiadou. 2002. "Automatic Discovery of Term Similarities Using Pattern Mining." In *Association for Computational Linguistics: the 19th International Conference on Computational Linguistics, COLING 2002 (COMPUTERM 2002: the second international workshop on computational terminology), August 24- September 1, 2002, Taipei, Taiwan*, 14: 1-7.

Neshati, Mahmood and Leila Sharif Hassanabadi. 2007. "Taxonomy Construction Using Compound Similarity Measure." In *On the Move to Meaningful Internet Systems 2007: CoopIS, DOA, ODBASE, GADA, and IS, OTM Confederated International Conferences CoopIS, DOA, ODBASE, GADA, and IS 2007, November 25-30, 2007, Vilamoura, Portugal. Lecture Notes in Computer Science 4804:* 915-32.

Patwardhan, Siddharth and Ted Pedersen. 2006. "Using WordNet-Based Context Vectors to Estimate the Semantic Relatedness of Concepts." In *Association for Computational Linguistics: Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), April 3-7, 2006, Trento, Italy,* 17, no. 6: 1-8.

Qiu, Zhengqiang. 2006. *Chinese Semantic Research.* Hunan: Central South University Press.

Ran, Jie and Yu Sun. 2011. "Research of Word Similarity Computing in Semantic Retrieval." *Computer Technology and Development* 21, no. 4: 94-7.

Schulz S, Hahn U. 2000. "Morpheme-Based, Cross-Lingual Indexing for Medical Document Retrieval." *International Journal of Medical Informatics,* nos. 58/59: 87-99.

Sigman, Mariano and Guillermo A. Cecchi. 2002. "Global Organization of the WordNet Lexicon." *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 99, no. 3: 1742-7.

Smith, T. F. and M. S. Waterman. 1981. "Identification of Common Molecular Subsequences." *Journal of molecular biology* 147, no. 1: 195-7.

Šojat, Krešimir and Matea Srebačić. 2014. "Morphosemantic Relations between Verbs in Croatian WordNet." In *the Global WordNet Association: Proceedings of the Seventh Global WordNet Conference, January 25-29, 2014, Tartu, Estonia,* ed. H. Orav, C. Fellbaum and P. Vossen, 262-7.

Wang, Yulin and Zhongyi Wang. 2014. "Research on Fine-grained Semantic Co-word Analysis Method." *Library and Information Service* 58, no. 21: 73-80.

Wu, Qingqiang and Yajuan Zhao. 2008. "Research in the Weighted Co-Word Analysis Based on the Attributes of Articles." *Journal of the China Society for Scientific and Technical Information* 27, no. 1: 89-92.

Yang, Liner and Maosong Sun. 2015. "Improved Learning of Chinese Word Embeddings with Semantic Knowledge." In *the 14th China National Conference, CCL 2015 and Third International Symposium, NLP-NABD 2015, November 13-14 2015, Guangzhou, China. Lecture Notes in Computer Science* 9427: 15-25.

Zhang, Chengzhi. 2005. "A Model for Chinese String Similarity Based on Multi-Level Features." *Journal of the China Society for Scientific and Technical Information* 24, no. 6: 696-701.

Zhang, Juan, Jun Xie, Wanli Hou, Xiaochen Tu, Jing Xu, Fujian Song, Zhihong Wang and Zuxun Lu. 2012. "Mapping the Knowledge Structure of Research on Patient Adherence: Knowledge Domain Visualization Based Co-Word Analysis and Social Network Analysis." *PloS one* 7, no. 4: e34497.

Zhang, Wei, Yang Yu and Hongliang You. 2009. "Relation Identification between Conceptual Terms for Automatic Construction of Lexical Knowledge Database." New *Technology of Library and Information Service* 185, no. 11: 10-6.

Zhu, Yihua, Hanqing Hou and Yingting Sha. 2002. "A Comparison of Two Algorithms for Computer Recognition of Chinese Synonyms." *The Journal of the Library Science in China* 28, no. 4: 82-5.

Zieman, Yuri and Ricardo Salas. 2001. "Semantic Labeling - Unveiling the Main Components of Meaning of Free-Text." In *IEEE Computer Society: Proceedings of the 8th International String Processing and Information Retrieval (SPIRE 2001), November 13-15 2001, Laguna de San Rafael, Chile.* 228-35.