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>

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.

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; Sojat and Srebatić 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 “烯” (poly-), “乙” (di-) and “烯” (alkene) in “烯乙烯” (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 concept. 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 $q_j$, i.e., $[w_i \rightarrow q_1 = \{c_1, c_2, ..., c_j \} | c_j \in C]$. 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):

$$R(W_i, W_j) = \left[ R_{sim}(W_i, W_j) \cup R_{rel}(W_i, W_j) \right] = a \cdot f_{sim}(Q_i, Q_j) + \beta \cdot f_{rel}(Q_i, Q_j)$$

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 ($q_j$) 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-
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” (poly-) “烯” (alkene) 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

Figure 1. Three phases of the term association process.

Figure 2. Composition of chemical morpheme collection.
“K-value” of terms is given by the sum of the weighted entropy of various morphemes, defined as follows:

\[ k\_value(w_i) = -\sum_{t_i} a_i \cdot p(t_i) \cdot \log_2 p(t_i) \]

\[ i = 1, 2, \ldots \]

Here, \( p(t) \) 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 \( a_i \) 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:

\[ R_{sim}(w_i, w_j) = \alpha \cdot \text{match}(w_i, w_j) + \beta \cdot \text{len}_{ratio}(w_i, w_j) \]

\[ \alpha + \beta = 1; \alpha, \beta > 0 \]

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

### 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:

\[ R_{rel}(w_i, w_j) = \frac{\text{sc}(w_i, w_j)}{\text{sc}(w_i) + \text{sc}(w_j) - \text{sc}(w_i, w_j)} \]

where \( \xi \) 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 \( \text{Contextspec} \) to measure the specific morpheme content of the context, indicating the specialization level of scientific literature as follows:

\[ \text{Contextspec} = \text{Count}(c_k) \]

where \( \text{Count}(c_k) \) measures the number of all \( k \)-type morphemes.
Context_spec\( (w_i, w_j) = \text{avg}(\text{count}(c_k)/\text{len}(s_{ij}^k)) \)

Here, \( s_{ij}^k \) stands for the sentence \( k \) where term \( w_i \) and \( w_j \) appear together, and \( \epsilon \) is the specific morpheme used in sentence \( k \). \( \text{Count}(c_j) \), \( \text{len}(w_i) \), separately measure the number of all specific morphemes in sentence \( k \) and the length of sentence \( k \). The specialization level index \( \text{ContextSpec} \) 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 \( \text{ContextSim} \) between terms \( w_i \) and \( w_j \) is the average similarity of any two morpheme sequences of context described as follows:

\[
\text{ContextSim}(w_i, w_j) = \text{avg}(\text{sim}(s_{ij}, s_{ij}'))
\]

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 \( \text{sim}(s, s') \) 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 \( R_{rel}(w_i, w_j) \). The equation is as follows:

\[
R_{rel} \ (w_i, w_j) = R_{rel}(w_i, w_j) * \text{ContextSpec}
\]

\[
(w_i, w_j)/\text{ContextSim} \ (w_i, w_j)
\]

### 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” (ethanol) are similar in Chinese word form, but belong to diols and monohydric alcohols, respectively. Hence, “醇” (alcohol) and “乙醇” (glycols) were added into term system. “乙醇” (glycols) was connected with the narrower term “乙二醇” (ethylene glycol), and “醇” (ethyl alcohol) was set as the broader term of “乙醇” (glycols) and “乙醇” (ethylene glycol). 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 \( k_{value} \).

- **Step 3**: Calculate the similarity \( R_{sim} \) of each term \( w_i \) with other terms \( w_j \) which should satisfy the condition that \( k_{value}(w_i) > k_{value}(w_j) \), i.e., term \( w_i \) is narrower than \( w_j \). 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 \( w_j \), term \( w_i \) 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 \( \epsilon \). We considered \( \epsilon = 0.015 \) because preliminary experiments show that for \( \epsilon \)
> 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:

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

### 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.

<table>
<thead>
<tr>
<th>Number</th>
<th>$D_1(800)$</th>
<th>$D_2(200)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keywords</td>
<td>1447</td>
<td>847</td>
</tr>
<tr>
<td>Initial chemical terms</td>
<td>834</td>
<td>509</td>
</tr>
<tr>
<td>Supplementary chemical terms</td>
<td>163</td>
<td>120</td>
</tr>
<tr>
<td>Connections based on semantic similarity</td>
<td>950</td>
<td>594</td>
</tr>
<tr>
<td>Connections based on semantic relatedness</td>
<td>1238</td>
<td>829</td>
</tr>
</tbody>
</table>

*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 “(methyl alcohol)” and “(ethyl alcohol)” directly connecting to the core node. Moreover, the narrower terms of “(alcohols)” include “(glycols)” (ethylene glycol), “(propylene glycol)” and “(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 nontrelated, 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.

<table>
<thead>
<tr>
<th></th>
<th>$D_1$</th>
<th>$D_2$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>85.03%</td>
<td>81.57%</td>
<td>76.67%</td>
</tr>
<tr>
<td>$P_2$</td>
<td>73.23%</td>
<td>73.30%</td>
<td></td>
</tr>
<tr>
<td>$P$</td>
<td>76.74%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*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., “(phenethyl alcohol)” was connected with “(phenethyl alcohol)” by a high similarity value, and the matching algorithm of the semantic core was designed to correct the mistake by linking “(phenethyl alcohol)” to “(phenethyl 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-
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 D2 appeared in D1, and > 80% of the semantic similarity links in D2 also appeared in D1. The main reason for these results is that the term nodes in D1, but not in D2, 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 D2 to
Table 3. Overlap ratio of semantic associations.

<table>
<thead>
<tr>
<th></th>
<th>Total connections</th>
<th>Connections based on semantic similarity</th>
<th>Connections based on semantic relatedness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occur in D2</td>
<td>1423</td>
<td>594</td>
<td>829</td>
</tr>
<tr>
<td>Co-occur in D1</td>
<td>1330</td>
<td>503</td>
<td>827</td>
</tr>
<tr>
<td>Overlap Ratio</td>
<td>93.46%</td>
<td>84.68%</td>
<td>99.76%</td>
</tr>
</tbody>
</table>

19.54% in D1, 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


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


Li, Nan, Rongting Zheng, Jiuming Ji and Qingqing Teng. 2010. “Research on Chinese Chemical Name Recogni-


