*ICIC Express Letters*
*Part B: Applications*
Volume **8**, Number **2**, February **2017**

ICIC International ©2017 ISSN 2185-2766
pp. **357–364**

# A FRAMEWORK FOR WORD SENSE DISAMBIGUATION OF ENGLISH MODAL VERBS BY FORMAL CONCEPT ANALYSIS

Jianping Yu[1,*], Hongbo Li[1] and Wenxue Hong[2]

[1]School of Foreign Studies
[2]Institute of Electrical Engineering
Yanshan University
No. 438, West Hebei Avenue, Qinhuangdao 066004, P. R. China
[*]Corresponding author: yjp@ysu.edu.cn

Received July 2016; accepted October 2016

Abstract. *Word sense disambiguation (WSD) is a hot and tough issue in machine learning and natural language processing, and English modal verbs are the semantically complex words whose senses are difficult to determine. Therefore, this paper puts forward a framework for the WSD of English modal verbs. The framework consists of five layers. The first layer is the linguistic preparation, in which the semantic model and feature system of English modal verbs are built up. Data collecting and data pre-processing are the following two preparatory processes. The fourth layer, the core process, is to establish WSD model by means of generating formal context and structural partial-ordered attribute diagram. The final layer is the model validation, in which the testing-data-set-validation method and N-fold-cross-validation method are used. The framework proposed in this study not only provides systematic methodological guidance for WSD of English modal verbs, but also sheds light on the WSD of words from other parts of speech.*
**Keywords:** Word sense disambiguation, Semantic model, Feature system, Formal context, Structural partial-ordered attribute diagram

1. **Introduction.** With the rapid development of Internet technology and its increasing popularity massive information floods to the Internet, the number of the texts in natural language processing (NLP) rises to an unprecedented level. To find a simple and accurate approach to word sense disambiguation (WSD) of natural language has become an urgent problem in NLP, human-computer interaction, artificial intelligence and other disciplines. Modal verbs are semantically complex words, and their semantic complexity makes it a great difficulty to determine their semantic meaning. Therefore, it is of great significance to find an effective way to the WSD of English modal verbs for NLP.

WSD mainly refers to the automatic recognition of the meaning of a word in a certain context. WSD is the basis of NLP research, and overlaps with information retrieval, machine translation, text classification and other areas. It has been a hot and difficult issue in NLP. There are three main approaches of WSD: dictionary based approach, instance based approach and corpus based approach. Dictionary based WSD method was proposed by Lesk [1] in 1986 which calculates the matching degree between the current text where the polysomy appears and the texts in which the meanings of words are interpreted in the dictionary, and selects the meaning in the maximum-matching-degree text as the correct meaning; the instances based method for WSD begins with example based machine translation, and the central idea of this method is to obtain WSD examples and calculate similarity among the examples; corpus based WSD method is to use statistical techniques to obtain the required knowledge automatically from the corpus for WSD. At present, the WSD studies conducted by domestic scholars also focus on these three aspects.

However, the author finds most of the WSD researches mainly focus on nouns, verbs, adjectives and some language structures, and the WSD studies on complex semantic words

such as English modal verbs are rare. Only Yu and her team have some contributions in this field [2-6]; in addition, most studies of WSD stop at the classification and clustering level and lack a summary of the WSD framework. The above two aspects indicate the study space for the WSD framework study of English modal verbs.

The contributions of the work are as follows. 1) It extends the part of speech of WSD from nouns, verbs and adjectives to modal verbs. 2) It provides a framework for the WSD of semantically complex words, such as English modal verbs.

The rest of the paper is organized as follows. Section 2 focuses on the theoretical background of the study. Section 3 presents a framework for the WSD of English modal verbs. Section 4 decomposes the framework and discusses each part in detail. Section 5 is the conclusion of the study.

## 2. Theoretical Background.

Formal concept analysis (FCA), a branch of mathematics, is proposed by Wille, a German professor, in 1982. FCA is different from traditional statistic data analysis and knowledge representation in that it is based on the formal concept and their hierarchy, and it studies attributes, objects and their interrelationships in the formal context. As an effective tool in data analysis and knowledge representation, FCA has developed rapidly in recent decades. The following definitions are central to FCA and this study [7].

**Definition 2.1.** *A formal context $K = (G, M, I)$ consists of two sets $G$ and $M$ and a relation $I$ between $G$ and $M$. The elements of $G$ are called the objects and the elements of $M$ are called the attributes of the context. In order to express that an object $g$ is in a relation $I$ with an attribute $m$, we write $gIm$ or $(g, m) \in I$ and read it as the object $g$ has attribute $m$.*

**Definition 2.2.** *Let $K = (G, M, I)$ be a formal context, for a set $A \subseteq G$, $f(A) = \{m \in M | (g, m) \in I, \forall g \in A\}$. Correspondingly, for a set $B \subseteq M$, we define $g(B) = \{g \in G | (g, m) \in I, \forall m \in B\}$. A formal concept is a pair $(A, B)$ with $A \subseteq G$, $B \subseteq M$, $f(A) = B$ and $g(B) = A$. $A$ is called extent and $B$ is called intent of the concept.*

**Definition 2.3.** *Let $K = (G, M, I)$ be a formal context, if for any objects $g_1, g_2 \in G$ from $f(g_1) = f(g_2)$, it always follows that $g_1 = g_2$ and correspondingly, $g(m_1) = g(m_2)$ implies $m_1 = m_2$ for all $m_1, m_2 \in M$, and the context $K = (G, M, I)$ is called clarified.*

## 3. A Framework for WSD of English Modal Verbs.

Yu et al. [8-11] have conducted some pioneer work in the field of WSD of semantically complex words including English modal verbs and have made great breakthrough in the study of WSD of semantically complex words. Based on those contributions, this study makes a summarization of the previous studies and proposes a framework (see Figure 1) which may be effective for the WSD of all English modal verbs.

## 4. Description of Each Layer in the Framework.

### 4.1. Linguistic preparation.

4.1.1. *Semantic model of English modal verbs.* As a complex semantic system, English modal verbs have the characteristics of ambiguity, merger and gradience, which cause great difficulty to both language learning and natural language processing. With reference to the relevant linguistic theories and some previous research findings [9], this paper summarizes a model (see Figure 2 and Figure 3) for the complex semantic system of English modal verbs, aiming at providing important guidance for the semantic studies of modal verbs in the following steps.

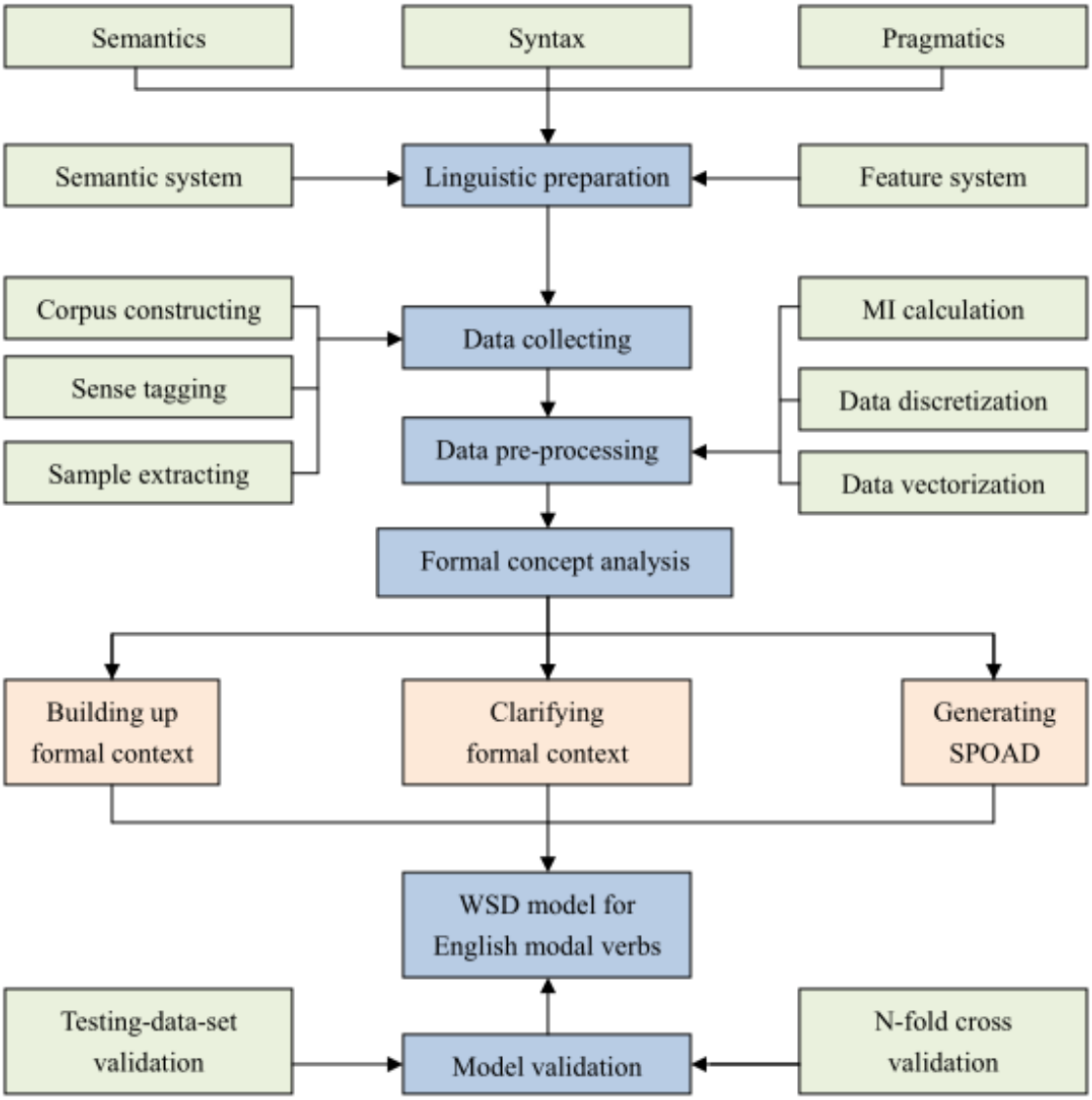


FIGURE 1. A framework for WSD of English modal verbs

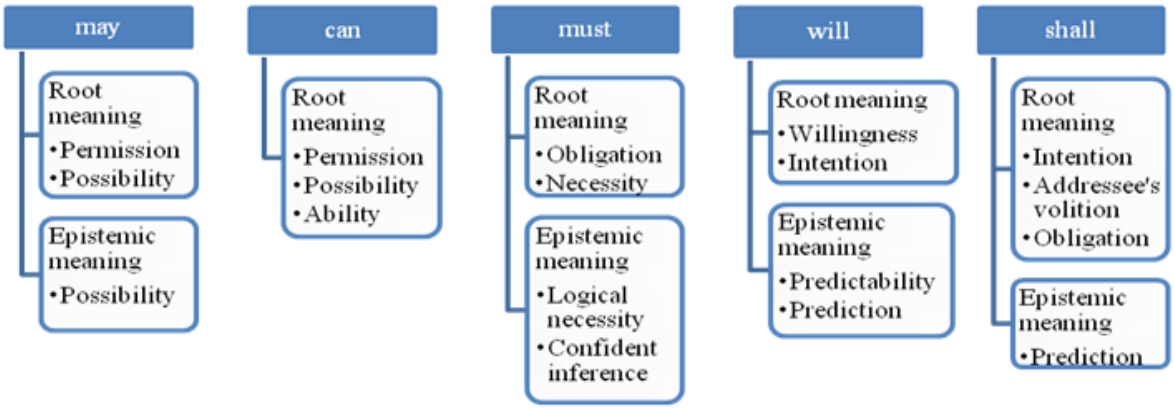


FIGURE 2. Semantic system for primary modal verbs

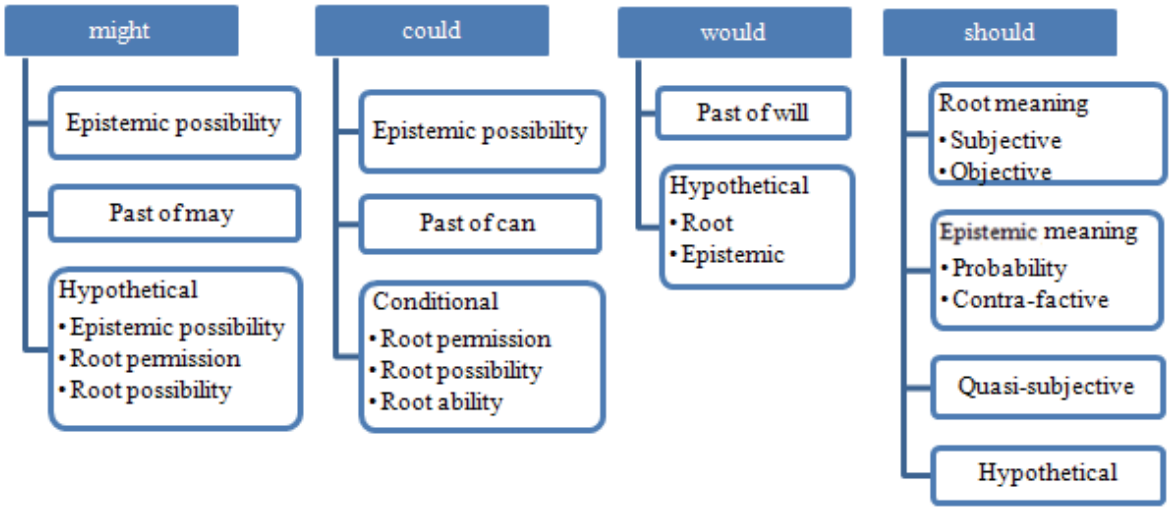


FIGURE 3. Semantic system for secondary modal verbs

4.1.2. *Feature system of English modal verbs.* The semantic complexity of modal verbs is influenced by different contexts from multiple dimensions and levels. Through observation and statistics of large-scale corpus and with the knowledge of semantics, syntax and pragmatics, a model for contextual features of modal verbs is constructed, which paves the way for the feature selection in WSD (see Figure 4).

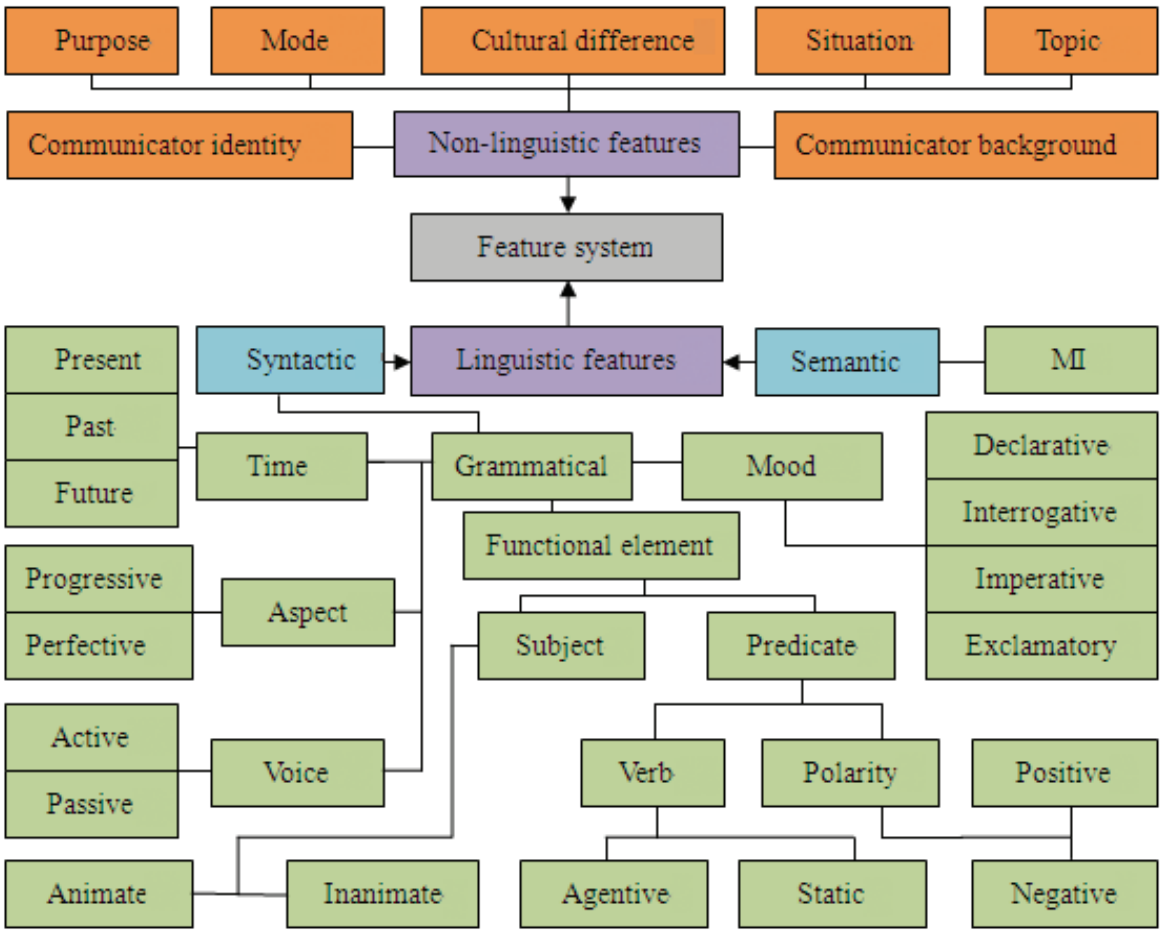


FIGURE 4. Feature system of English modal verbs

According to this model, the context features can be classified into linguistic context features and nonlinguistic context features, and the two categories include many intertwined and interacted aspects. Linguistic features are divided into semantic features and syntactic features. Semantic features are symbolized by mutual information (MI) between modal verbs and the adjacent subjects and the main verbs; syntactic features include three categories: grammatical, functional elements and mood of sentences. Grammatical elements consist of time (present, past and future), aspect (progressive and perfective) and voice (active and passive), functional elements include subject (animate and inanimate) and predicate (verb and polarity), and the mood of sentences covers declarative sentence, interrogative sentence, imperative sentences and exclamatory sentence. Nonlinguistic context mainly refers to the non-language factors that affect the language, such as communicator identity, communicator background, purpose, mode, cultural difference, situation and topic.

4.2. **Data collecting.** Data collecting in this study includes three steps: corpus constructing, sense tagging and sample extracting.

4.2.1. *Corpus constructing.* This study tries to study the English modal verbs in different types of current human language, which means the source of the corpora must be various, credible and fresh. According to this principle, a 1.8-million-word corpus is established. The genres of the materials include research articles, novels, news reports, legal documents, public speeches, interviews and movie lines, each accounting for 0.25 million words in order to maintain the balance of different genres. It can also be seen that the channels of the materials vary from written to spoken.

4.2.2. *Sense tagging.* First, determine the tagging set and tagging standard for each English modal verb. According to the semantic system of English modal verbs, the senses of each modal verb are manually tagged. Then, by means of cross examination among annotators and experts, a corpus with the semantic annotation of English modal verbs is formed.

4.2.3. *Sample extracting.* At this step, the prepared corpus is changed into different data sets. For each English modal verb, all the sentences containing different senses of it will be copied into a new text and those sentences will be divided into different groups according to different senses. Then, according to the random sampling principle, sample sentences are chosen for each sense. If there are enough sentences, a training data set and two testing data sets will be set up; if not, all the sentences are kept in the data set.

4.3. **Data pre-processing.**

4.3.1. *Mutual information calculation.* Mutual information (MI), which expresses the semantic correlation between two words, is considered as semantic features in this study. MIs between subject and *must*, *must* and the following verb in each sample sentence are calculated according to Formula (1):

$$MI(w_1, w_2) = \log \frac{P(w_1, w_2)}{P(w_1)P(w_2)} \tag{1}$$

Here, $w_1$ and $w_2$ are two words. In this study, $w_1$ is one specific English modal verb and $w_2$ is the subject or the main verb in the sample sentence. $P(w_1)$ and $P(w_2)$ represent the probabilities of $w_1$ and $w_2$ that appear independently in the whole corpus, while $P(w_1, w_2)$ stands for the probability of the co-occurrence of $w_1$ and $w_2$ in the whole corpus. The number of the pairs of MIs to be calculated depends on how many senses the specific English modal verb possesses: MI(s, $\text{sense}_1$), MI(s, $\text{sense}_2$), ..., MI(s, $\text{sense}_n$); MI($\text{sense}_1$,v), MI($\text{sense}_2$,v), ..., MI($\text{sense}_n$,v). MI(s, $\text{sense}_n$) is the mutual information

between subject and the sense category $n$ of the modal verb; MI(sense$_n$, v) is the mutual information between the sense category $n$ of the modal verb and the main verb.

4.3.2. *Data discretization.* All the MIs obtained from Section 4.3.1 are continuous values; however, the approach used in this study cannot directly deal with continuous variables, and thus all the MIs need discretization. We have used two methods to discrete MIs: one is equal interval division; the other is scatter diagram division. The former means to select dividing points evenly among the MIs' values, for example, MI $\leq -0.5$, $-0.5 <$ MI $\leq 0$, $0 <$ MI $\leq 0.5$, MI $> 0.5$ (see [8]); the latter approach is to generate a scatter diagram for each group of MI, then it can be observed from the diagram that at which points the different senses are best discriminated, so the divided ranges are obtained (see [2,3]).

## 4.4. **Formal concept analysis.**

4.4.1. *Building up the formal context.* Objects and attributes are the two basic elements in a formal context. In this study, objects are the sample sentences with different senses of the English modal verb and attributes are the contextual features of it. Since formal concept analysis can only process binary values, all the features should be vectorized into binary values. According to the values and the dividing ranges of MIs, we first symbolize different features of MIs with 1 or nothing. If the value of certain MI falls into certain range as a$_n$, then this blank is marked as 1; otherwise, nothing is given to it. The syntactic features and non-linguistic features are dealt with in the same way according to the presence or absence of the feature. If the sample co-occurs with the feature, a logical value of 1 is given to it; otherwise, nothing. Based on this symbolization, the formal context of the English modal verb can be obtained (see Table 1 as an example).

Table 1. A formal context

| *ai* / *oj(s)* | a1 | a2 | a3 | a4 | a5 | a6 | a7 | a8 | a9 |
|---|---|---|---|---|---|---|---|---|---|
| o1(1) | | | 1 | | | | | | |
| o2(1) | | | 1 | | | | | | |
| o3(1) | | | 1 | | | | | | |
| o4(1) | 1 | | | | | | | | |
| o5(1) | | | | | | | | 1 | 1 |
| o6(1) | | | | | | | | | |
| o7(1) | | | | | | | | 1 | 1 |
| o8(1) | | | | | | | | | |
| o9(1) | | 1 | | | | | | | 1 |
| o10(1) | | | | | | | | 1 | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| o91(2) | | | | | | | | 1 | 1 |
| o92(2) | | | 1 | | | 1 | | | |
| o93(2) | | | | | 1 | 1 | | | |
| o94(2) | | | | | 1 | | | | |
| o95(2) | | | | | | 1 | | | |
| o96(2) | | | | | | | | 1 | |
| o97(2) | | | | | 1 | | | | |
| o98(2) | | | | | 1 | | | 1 | |
| o99(2) | | | | | 1 | | | 1 | |
| o100(2) | | | | | 1 | | | 1 | |

4.4.2. *Clarifying the formal context.* According to Definition 2.3, in the original formal context, only one of the objects sharing the same attributes is reserved and the others are deleted, so the clarified formal context is built.

4.4.3. *Generating SPOAD.* The SPOAD tool [12] is used to convert the formal context into a corresponding hierarchical relation diagram. SPOAD construction is based on the theory of formal concept analysis (FCA). The principle of the approach is as follows: in the clarified formal context, all concepts will be classified according to the attributes. A super-class must have the common attributes of subclasses, and a subclass either inherits from a super-class, or is an independent attribute of class elements. If the attributes in the same layer have intersections, there must be corresponding sub-nodes at the lower layer. According to this principle, we start with the top node to build the attribute hierarchical diagram, generate all the sub-nodes of the top node, and then repeat the same process for each sub-node until no sub-node can be classified. Since the features clustered in each line in SPOAD form a pattern to realize the sense classification of the English modal verbs, this diagram can function as a WSD model for the English modal verb (See Figure 5 as an example).

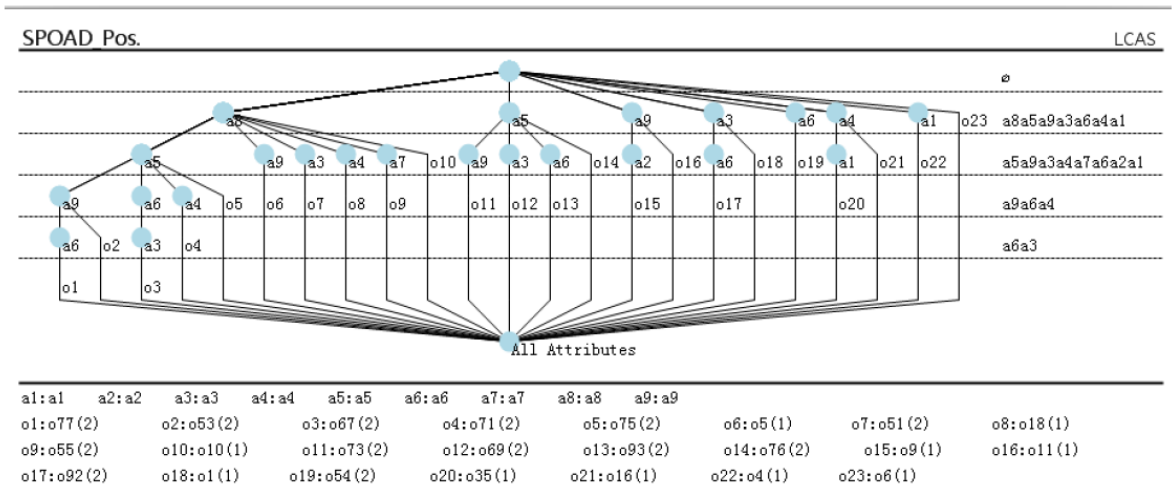


FIGURE 5. The clarified SPOAD of the formal context in Table 1

## 4.5. **Testing of the WSD model.**

4.5.1. *Testing-data-set validation.* If there are enough samples of the studied English modal verb, the two testing data sets mentioned in Section 4.2.3 are used to test the accuracy of the WSD model. The data sets are processed with the same procedure mentioned in Section 4.3 and Section 4.4.1, and thus we obtain the formal context of the testing data sets. Then each object is examined with the pattern in the WSD model. For the object possessing exactly the same feature cluster as the one in the model, it is certain that it belongs to the same sense group as the one in the model; and for the ones not having the same feature cluster, the similarity principle is used. If the similarity is equal or greater than 2/3, it belongs to the same sense group as the one in the model. The mean value of the accuracy rates of the two testing data sets is the final result of the validation.

4.5.2. *N-fold cross validation.* If there is only one data set in the experiment and there are no testing data sets to verify the accuracy, the method of $N$-fold cross validation should be adopted to test the effectiveness of the model. The procedure is as follows. First, the original data set is divided evenly into $N$ groups. Then, each group is chosen to be the testing set and the other $N-1$ groups are used as the training set to carry out

the experiment. The accuracy of these experiments can be calculated. The average value of the experiments accuracy plus their standard deviation is the result of $N$-fold cross validation for the WSD model.

5. **Conclusion.** This paper proposes a framework for WSD of English modal verbs. A prerequisite for WSD is to establish a semantic model and a feature system model for English modal verbs. Then data collecting and data pre-processing are the two preparatory steps to get the crucial data for the experiment. The core process is the WSD model construction by generating the formal context and the SPOAD for English modal verbs. The final layer in the framework is the WSD model validation. The paper provides systematic methodological guidance for the WSD of English modal verbs and it can also be a sample for the WSD of the words from other parts of speech.

**Acknowledgements.** This work is supported by the Social Sciences Foundation of Hebei Province under Grant No. HB15YY02. It is also supported by the Humanities and Social Sciences Foundation of the Ministry of Education of China under Grant No. 14YJC740038 and the Social Sciences Foundation of Hebei Province under Grant No. HB14YY005. The authors gratefully acknowledge the supports.

## REFERENCES

[1] M. Lesk, Automated sense disambiguation using machine-readable dictionaries: How to tell pine cone from an ice cream cone, *Proc. of the 1986 SIGDOC Conference*, Toronto, Canada, pp.24-26, 1986.

[2] J. Yu, W. Hong, C. Qiu, S. Li and D. Mei, A new approach of attribute partial order structure diagram for word sense disambiguation of English prepositions, *Knowledge-Based Systems*, vol.95, pp.142-152, 2016.

[3] J. Yu, C. Li, W. Hong, S. Li and D. Mei, A new approach of rules extraction for word sense disambiguation by features of attributes, *Applied Soft Computing Journal*, vol.27, pp.411-419, 2015.

[4] J. Fu, J. Yu and H. Liu, An investigation of influence of different subjective factors to WSD of English modal verb CAN, *ICIC Express Letters, Part B: Applications*, vol.6, no.5, pp.1473-1478, 2015.

[5] X. Xu, J. Yu and X. Piao, Contribution of governors to word sense disambiguation of English preposition, *ICIC Express Letters, Part B: Applications*, vol.6, no.3, pp.723-730, 2015.

[6] H. Li and J. Yu, Attribute significance analysis of English modal verb shall in word sense disambiguation, *ICIC Express Letters, Part B: Applications*, vol.6, no.5, pp.1287-1294, 2015.

[7] B. Ganter and R. Wille, *Formal Concept Analysis*, Spring-Verlag, Berlin, 1999.

[8] J. Yu, N. Chen, R. Sun, W. Hong and S. Li, Word sense disambiguation and knowledge discovery of English modal verb can, *ICIC Express Letters*, vol.7, no.2, pp.577-582, 2013.

[9] J. Yu, W. Hong, S. Li, T. Zhang and J. Song, A new approach of word sense disambiguation and knowledge discovery of English modal verbs by formal concept analysis, *International Journal of Innovative Computing, Information and Control*, vol.9, no.3, pp.1189-1200, 2013.

[10] J. Yu, W. Hong, S. Zhang and S. Zhao, Improving the precision of word sense disambiguation of English modal verb will by formal concept analysis, *ICIC Express Letters, Part B: Applications*, vol.3, no.4, pp.751-757, 2012.

[11] J. Yu, *Intelligent Word Sense Disambiguation of English Modal Verbs*, Shanghai International Studies University, 2011.

[12] W. Hong, S. Li, J. Yu and J. Song, A new approach of generation of structural partial ordered attribute diagram, *ICIC Express Letters, Part B: Applications*, vol.3, no.4, pp.823-830, 2012.