ICIC Express Letters ICIC International ©2016 ISSN 1881-803X
Volume **10**, Number **6**, June **2016** pp. 1325–1333

A RULE-EXTRACTION-BASED OPTIMIZATION METHOD FOR FEATURE SELECTION IN WORD SENSE DISAMBIGUATION

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

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

Received November 2015; accepted February 2016

Abstract. *Feature selection is an important process in classification and pattern recognition and it has a direct influence on the accuracy of classifier. In this study, a new optimization method of feature selection by means of rule extraction is proposed for word sense disambiguation (WSD) of English modal verb "must". A WSD model with all candidate features for "must" is constructed first with the approach of structural partial-ordered attribute diagram (SPOAD) and the accuracy of WSD is tested to be 94.5%. Then based on the WSD model and the rule-extraction algorithm, rules for the two senses of "must" are extracted, and accordingly the optimized feature set with only 6 attributes is obtained. The WSD model with the optimized feature set yields a classification accuracy of 97.5%, which is 3% higher than that of the original model. Therefore, it is concluded that the proposed method can optimize the feature set and is effective in dealing with binary classification problems in WSD. It can also be applied to other binary-classifier research and provides valuable reference for feature selection in machine learning and natural language processing.*

Keywords: Feature selection, Rule extraction, Semantic features, Syntactic features, Structural partial-ordered attribute diagram

1. **Introduction.** Feature selection, also called feature subset selection or attribute selection, refers to the process of selecting an optimal subset from all the candidate features in order to improve the classifier performance. In the field of machine learning or pattern recognition, features have expanded from tens to hundreds of variables, increasing the computational time and the complexity of the model and decreasing the performance precision and the generalization ability of the model. Several techniques are developed to address the problem by reducing irrelevant and redundant features which are a burden for the classifier, namely, feature selection. Feature selection helps in understanding data, reducing computation requirement, avoiding the dimension disaster and improving the classifier performance [1].

Feature selection has been the focus of interest for quite some time and much work has been done. As defined by John and Kohavi [2], this work is broadly divided into two categories: filter and wrapper methods. In the filter methods, feature selection acts as a preprocessing step to rank the features and the highly ranked features are selected and applied to a model. Two of the most well-known ranking criteria for filter methods are correlation criteria [3,4] and Mutual Information [5,6]. The former detects the linear dependencies between feature and concept, while the latter uses the measure of dependency between two features. The filter methods are computationally light, but the drawback is that the selected subset might not be optimal in that a redundant subset might be obtained.

In wrapper methods, the model searches through the space of feature subsets using the accuracy from a learning algorithm as the measure of goodness of particular feature subset, i.e., the learning algorithm is wrapped on a search algorithm which will find a subset which outputs the highest model performance. The wrapper methods are generally classified into Sequential Selection Algorithms [7,8] and Heuristic Search Algorithms [9,10]. Using wrapper methods is likely to produce a feature subset with high accuracy but the number of computation is large.

Both filter methods and wrapper methods for feature selection are proven to work well for certain datasets, and the present literature related to feature selection has made a great contribution to machine learning and natural language processing. However, few studies of feature selection have been conducted from the perspective of word sense disambiguation (WSD) and no studies have reported a rule-extraction-based feature selection method.

Word sense disambiguation is a hot and tough issue in machine learning and natural language processing. Since many contextual features coexist with the target word, it is of vital importance to select the optimal feature subset which can disambiguate the target word with less effort and higher accuracy. In light of this, we design this study on one case of English modal verb *must* (for more modal verb WSD studies, refer to [11-13]) and propose a new rule-extraction-based feature selection method for its WSD. The feature selection method proposed in this article can also be applied to the WSD of other words, and the study can provide some insight for machine learning and natural language processing.

The rest of the paper is organized as follows. Section 2 focuses on data collecting and pre-processing. Section 3 demonstrates how a WSD model with all candidate features of *must* is constructed. Section 4 presents the algorithm for the optimization of feature selection in this experiment and discusses its effectiveness. Section 5 is the conclusion of the study.

2. **Data Collecting and Pre-Processing.**

2.1. **Data collecting.** In this study, *must* is taken as the target word for feature selection in WSD. *Must* is one of the English primary modal verbs with high frequency [14] and the study of its feature selection and disambiguation is important for the studies of other modal verbs as well as for natural language processing tasks. *Must* is mainly categorized into two senses [15]: a root meaning and an epistemic meaning. Examples of each sense are shown in Table 1:

Table 1. Sense classification of *must*

Sense of *must*	Examples
Root	"You *must* play this ten times over," said Miss Jarrova.
Epistemic	You *must* have thought about that.

The ambiguity of *must* may cause trouble in machine learning and natural language processing, so it is of urgent need to carry out a WSD study of *must*, among which feature selection is a crucial step. Data collecting in this study includes three steps: corpus constructing, sense tagging and sample extracting. A 1.8-million-word corpus is prepared for the study, including novels, news reports, research articles, legal documents, public speeches, interviews and movie lines. Then, according to the above-mentioned classification standard, the senses of *must* are manually tagged. In total, the original corpus provides 750 sense-tagged *must*-sentences, among which 505 instances donate root *must* (RT*must*) and 245 instances donate epistemic *must* (EPI*must*). Finally, 200 RT*must* sample sentences and 200 EPI*must* sample sentences are extracted from the corpus to build up a training dataset and a testing dataset, each including 100 RT*must* and 100 EPI*must*.

2.2. Data pre-processing.

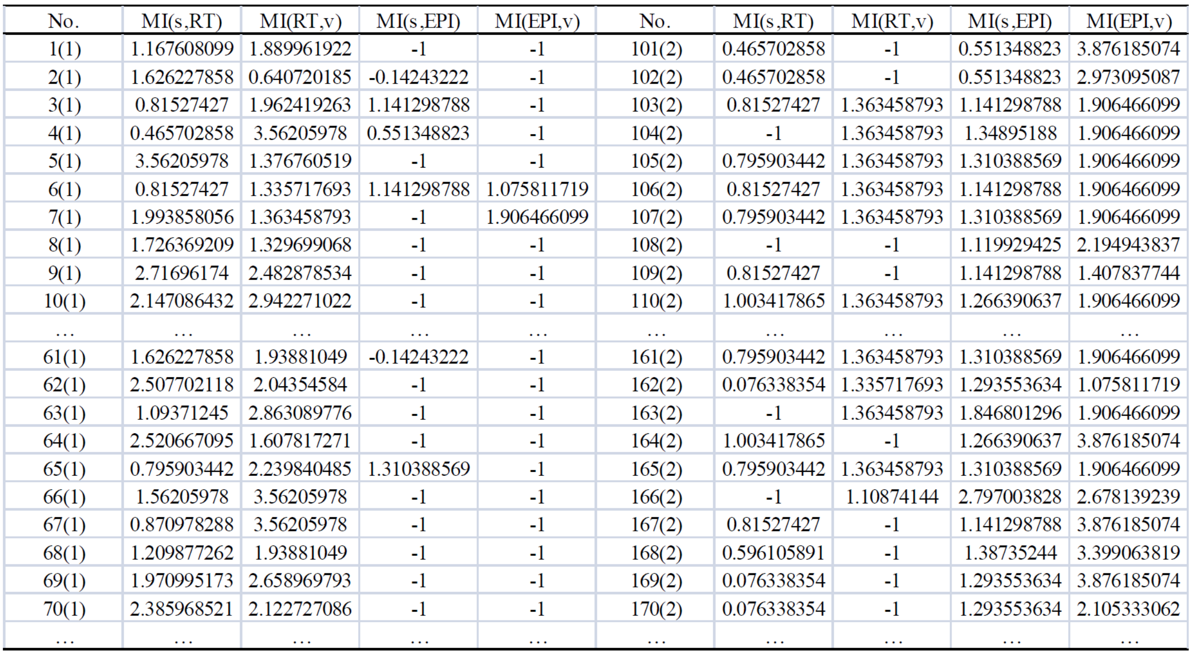
2.2.1. *Mutual information calculation.* Mutual information (MI), which expresses the semantic correlation between two words, is considered as semantic features in this study. MIs between subject and *must*, *must* and the following verb in each sample sentence are calculated according to Formula (1):

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

Here, w_1 and w_2 are two words. In this study, w_1 is *must* 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. Since there are two senses for *must*, there are four pairs of MIs to be calculated: MI(s, RT*must*), MI(RT*must*, v), MI(s, EPI*must*) and MI(EPI*must*, v). MI(s, RT*must*) is the mutual information between subject and RT*must*, MI(RT*must*, v) is the mutual information between RT*must* and the main verb, MI(s, EPI*must*) is the mutual information between subject and EPI*must*, and MI(EPI*must*, v) is the mutual information between EPI*must* and the main verb.

In calculating MIs, it is found when the co-occurrence frequency of two words is 0, the value of MI is symbolized by $\#NUM$, which represents negative infinity. In such cases, a value should be assigned to $\#NUM$; otherwise, it defaults to be 0 during the following step. So $\#NUM$ is replaced by -1, which is smaller than the minimum value in all MIs. In this way, all the MIs are obtained, as shown in Table 2.

Table 2. Results of MI calculation



No.	MI(s,RT)	MI(RT,v)	MI(s,EPI)	MI(EPI,v)	No.	MI(s,RT)	MI(RT,v)	MI(s,EPI)	MI(EPI,v)
1(1)	1.167608099	1.889961922	-1	-1	101(2)	0.465702858	-1	0.551348823	3.876185074
2(1)	1.626227858	0.640720185	-0.14243222	-1	102(2)	0.465702858	-1	0.551348823	2.973095087
3(1)	0.81527427	1.962419263	1.141298788	-1	103(2)	0.81527427	1.363458793	1.141298788	1.906466099
4(1)	0.465702858	3.56205978	0.551348823	-1	104(2)	-1	1.363458793	1.34895188	1.906466099
5(1)	3.56205978	1.376760519	-1	-1	105(2)	0.795903442	1.363458793	1.310388569	1.906466099
6(1)	0.81527427	1.335717693	1.141298788	1.075811719	106(2)	0.81527427	1.363458793	1.141298788	1.906466099
7(1)	1.993858056	1.363458793	-1	1.906466099	107(2)	0.795903442	1.363458793	1.310388569	1.906466099
8(1)	1.726369209	1.329699068	-1	-1	108(2)	-1	-1	1.119929425	2.194943837
9(1)	2.71696174	2.482878534	-1	-1	109(2)	0.81527427	-1	1.141298788	1.407837744
10(1)	2.147086432	2.942271022	-1	-1	110(2)	1.003417865	1.363458793	1.266390637	1.906466099
…	…	…	…	…	…	…	…	…	…
61(1)	1.626227858	1.93881049	-0.14243222	-1	161(2)	0.795903442	1.363458793	1.310388569	1.906466099
62(1)	2.507702118	2.04354584	-1	-1	162(2)	0.076338354	1.335717693	1.293553634	1.075811719
63(1)	1.09371245	2.863089776	-1	-1	163(2)	-1	1.363458793	1.846801296	1.906466099
64(1)	2.520667095	1.607817271	-1	-1	164(2)	1.003417865	-1	1.266390637	3.876185074
65(1)	0.795903442	2.239840485	1.310388569	-1	165(2)	0.795903442	1.363458793	1.310388569	1.906466099
66(1)	1.56205978	3.56205978	-1	-1	166(2)	-1	1.10874144	2.797003828	2.678139239
67(1)	0.870978288	3.56205978	-1	-1	167(2)	0.81527427	-1	1.141298788	3.876185074
68(1)	1.209877262	1.93881049	-1	-1	168(2)	0.596105891	-1	1.38735244	3.399063819
69(1)	1.970995173	2.658969793	-1	-1	169(2)	0.076338354	-1	1.293553634	3.876185074
70(1)	2.385968521	2.122727086	-1	-1	170(2)	0.076338354	-1	1.293553634	2.105333062
…	…	…	…	…	…	…	…	…	…

2.2.2. *Data discretization.* Because the algorithm in this study cannot directly deal with continuous variables, all the MIs need discrete treatment. To achieve this, a scatter diagram is generated for each group of MI, take the MI(s, RT*must*) for example, and see Figure 1, in which dots of different shapes stand for MIs between subject and RT*must*, with rhombus dots representing MIs of the first 100 sample sentences with RT*must* and square dots representing the other 100 sentences with EPI*must*. It can be seen for the figure that the two groups are best discriminated around the value 1.01, and thus the dividing ranges are set to be $a_1 \leq 1.01$; $a_2 > 1.01$. The dividing ranges of the other MIs are decided in the same way and altogether there are 8 ranges:

a_1: MI(s, RT*must*) ≤ 1.01
a_2: MI(s, RT*must*) > 1.01
a_3: MI(RT*must*, v) ≤ 1.40
a_4: MI(RT*must*, v) > 1.40
a_5: MI(s, EPI*must*) ≤ 0.15
a_6: MI(s, EPI*must*) > 0.15
a_7: MI(EPI*must*, v) ≤ 0.85
a_8: MI(EPI*must*, v) > 0.85

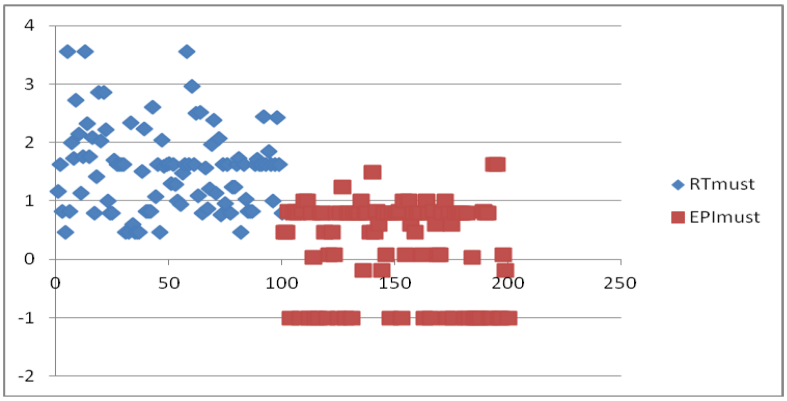


Figure 1. Scatter diagram of MI(s, RT*must*)

2.2.3. *Feature vectorization.* In addition to semantic features, there are several syntactic co-occurrence features for *must*: 1) negation; 2) passive; 3) second person subject; 4) first person subject; 5) perfective aspect; 6) progressive aspect; 7) existential subject; 8) stative verb; 9) inanimate subjects. Till now, all the candidate features are selected to construct the formal context and the SPOAD of *must*.

SPOAD construction is based on the theory of formal concept analysis (FCA). FCA is a method mainly used for the analysis of data. The following definition is central to FCA and is used in the study [16]:

Definition 2.1. *U is the set of objects, $U = \{u_1, u_2, \ldots, u_n\}$. M is the set of attributes, $M = \{m_1, m_2, \ldots,\ m_n\}$ and $I \subseteq U \times M$ is a binary relation between U and M with $(u, m) \in I$ indicating that the object u owns the attribute m. Then, $K = (U, M, I)$ is named as a formal context.*

Objects and attributes are the two basic elements in a formal context. In this paper, objects are the sample sentences with different senses of *must* and attributes are the linguistic features. Since formal concept analysis can only process binary values, all the extracted features should be vectorized into binary values. According to the values and the dividing ranges of MIs, we first symbolize a_1-a_8 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 nine syntactic features a_9-a_{17} 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 with all candidate features of English modal verb *must* is obtained (see Table 3).

3. Construction of WSD Model with All Candidate Features.

In this section, we intend to disambiguate the word senses of *must* according to the features they have. For convenience of exhibition and the subsequent rule extraction, we first clarify the formal context with all candidate features in Table 3 according to Definition 3.1. Among the objects that share the same attributes, only one is reserved and the others are deleted; thus the clarified formal context is built.

Table 3. Formal context of *must*

No.	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a_{17}
o1(1)		1		1	1		1			1							1
o2(1)		1	1		1		1					1					
o3(1)	1			1		1	1				1						
o4(1)	1			1		1	1					1					
o5(1)		1	1		1		1										
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
o91(1)		1		1	1		1		1			1					
o92(1)		1		1	1		1										
o93(1)		1		1	1		1					1					
o94(1)		1		1	1		1									1	1
o95(1)		1		1	1		1					1					
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
o131(2)	1		1			1		1								1	1
o132(2)	1		1			1		1								1	1
o133(2)	1		1			1		1					1			1	1
o134(2)	1		1			1		1					1			1	1
o135(2)	1		1			1		1					1	1		1	
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
o196(2)	1		1			1		1									1
o197(2)	1		1			1		1									1
o198(2)	1		1			1		1					1			1	
o199(2)	1		1			1		1					1			1	1
o200(2)	1		1			1		1							1	1	

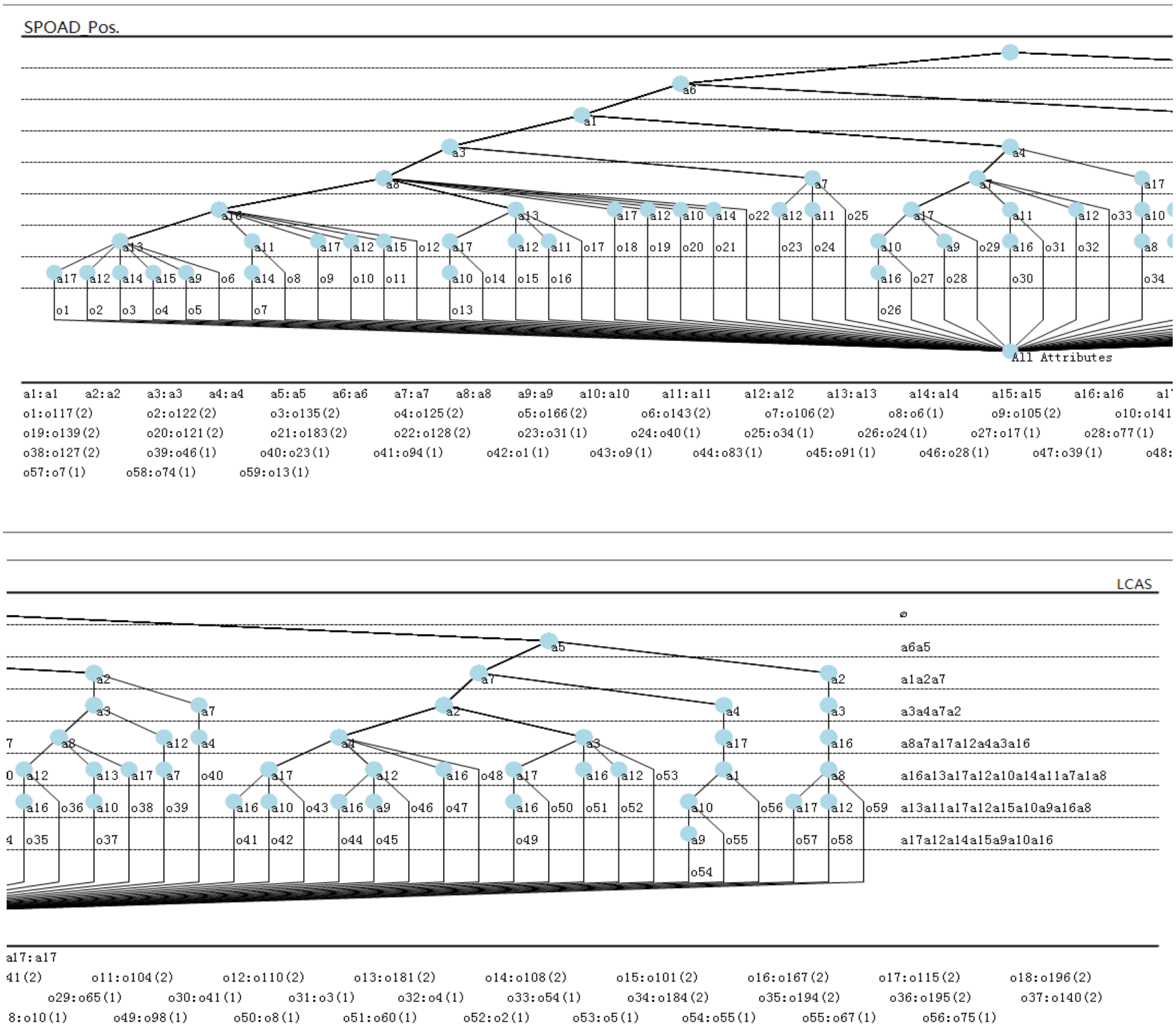

SPOAD_Pos.
All Attributes
a1:a1 a2:a2 a3:a3 a4:a4 a5:a5 a6:a6 a7:a7 a8:a8 a9:a9 a10:a10 a11:a11 a12:a12 a13:a13 a14:a14 a15:a15 a16:a16 a1
o1:o117(2) o2:o122(2) o3:o135(2) o4:o125(2) o5:o166(2) o6:o143(2) o7:o106(2) o8:o6(1) o9:o105(2) o10:o141
o19:o139(2) o20:o121(2) o21:o183(2) o22:o128(2) o23:o31(1) o24:o40(1) o25:o34(1) o26:o24(1) o27:o17(1) o28:o77(1)
o38:o127(2) o39:o46(1) o40:o23(1) o41:o94(1) o42:o1(1) o43:o9(1) o44:o83(1) o45:o91(1) o46:o28(1) o47:o39(1) o48:
o57:o7(1) o58:o74(1) o59:o13(1)
LCAS
∅
a6a5
a1a2a7
a3a4a7a2
a8a7a17a12a4a3a16
a16a13a17a12a10a14a11a7a1a8
a13a11a17a12a15a10a9a16a8
a17a12a14a15a9a10a16
a17:a17
41(2) o11:o104(2) o12:o110(2) o13:o181(2) o14:o108(2) o15:o101(2) o16:o167(2) o17:o115(2) o18:o196(2)
o29:o65(1) o30:o41(1) o31:o3(1) o32:o4(1) o33:o54(1) o34:o184(2) o35:o194(2) o36:o195(2) o37:o140(2)
8:o10(1) o49:o98(1) o50:o8(1) o51:o60(1) o52:o2(1) o53:o5(1) o54:o55(1) o55:o67(1) o56:o75(1)

Figure 2. The clarified SPOAD of *must*

Definition 3.1. *Let $K = (U, M, I)$ be a formal context, if for any objects $u_1, u_2 \in U$ from $f(u_1) = f(u_2)$, it always follows that $u_1 = u_2$ and correspondingly, $g(m_1) = g(m_2)$ implies $m_1 = m_2$ for all $m_1, m_2 \in M$, the context $K = (U, M, I)$ is called clarified.*

Then the SPOAD tool [17] is used to convert the formal context into a corresponding hierarchical relation diagram (see Figure 2). Since the feature cluster in each line is a pattern to realize the sense classification of English modal verb *must*, this diagram can function as WSD model for *must*.

To test the accuracy of the WSD model with all candidate features, the testing data set is processed with the same procedure mentioned in 2.2; thus we obtain the formal context of the testing data set. 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 should be examined whether 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 followed to see whether this object owns the same meaning as the objects having most of the features in the model. All the 200 objects in the testing data set are observed in this way and the accuracy is 94.5%.

4. **Optimization of Feature Selection.**

The WSD model for *must* with all candidate features is effective and the accuracy is high in disambiguating the two senses of *must*. In this part, we continue to investigate whether we can achieve the same effectiveness and accuracy for the WSD of *mus*t with fewer features. We propose a rule-extraction-based optimization method for this purpose.

4.1. **Rule extraction.**

Based on the SPOAD in Figure 2, the rules for WSD of *must* are extracted according to the rule-extraction flowchart (see Figure 3).

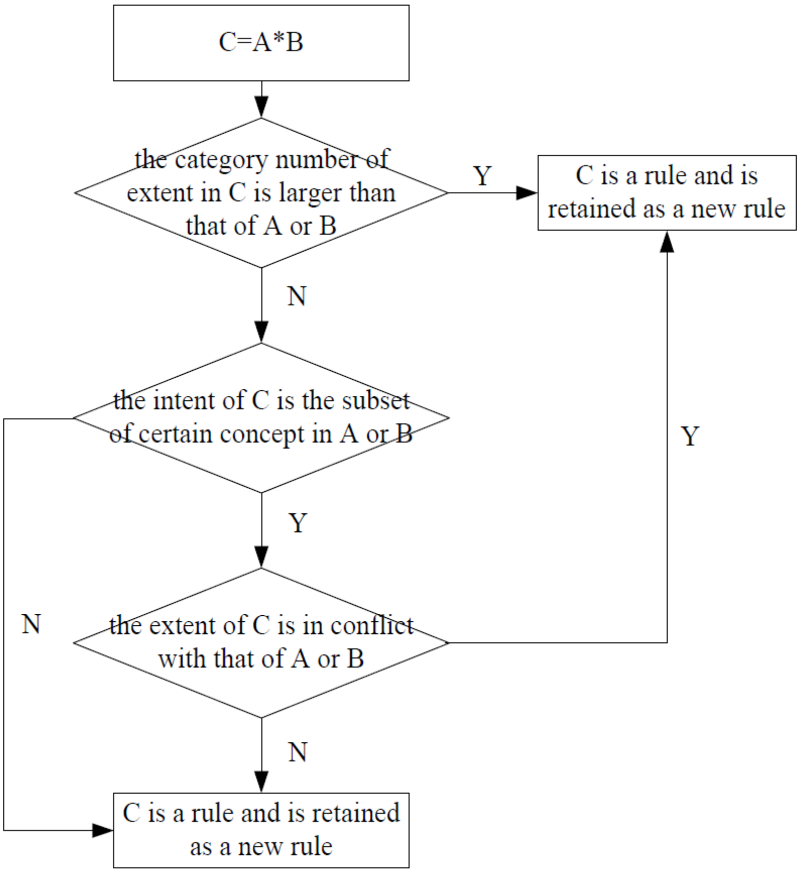

C=A*B
the category number of extent in C is larger than that of A or B
Y
C is a rule and is retained as a new rule
N
the intent of C is the subset of certain concept in A or B
Y
N
Y
the extent of C is in conflict with that of A or B
N
C is a rule and is retained as a new rule

FIGURE 3. Rule-extraction flowchart

Find a_i in the SPOAD, and then go up the lines to the top node; the attribute combination constitutes a in the concept (g, a); then go down the lines to the objects, and the object set constitutes g in the concept (g, a). After this operation, many concepts (the possible rules) may be obtained for a_i. All the possible rules are compared to observe whether they share the same intent. If there are some concepts sharing the same intent, a new pair will be generated with the same intent but the integrated extent; meanwhile, the original concepts are deleted. As for the extent, if two concepts share the same extent, the intent must be in the relation of inclusion, then the one with the largest intent (with the most attributes) is retained and the rest of the concepts are deleted.

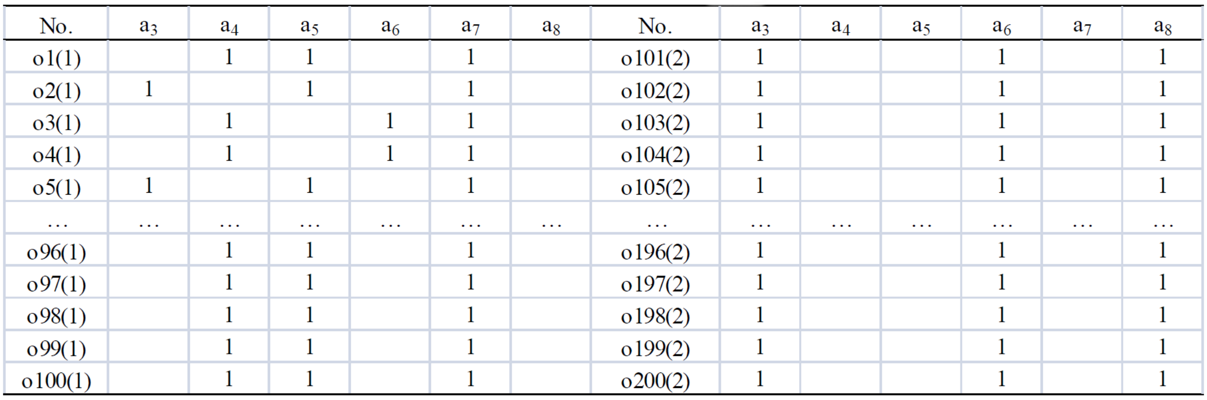
After this process, the concepts retained are the rules extracted from the SPOAD. In this study, the rules extracted from Figure 2 for each sense of *must* are listed in Table 4.

Table 4. Extracted rules of *must*

Rules for RT*must*	$\{1, a_4\}$	$\{1, a_5\}$	$\{1, a_7\}$
Rules for EPI*must*	$\{2, a_3a_6a_8\}$		

4.2. **Optimization of feature selection.** The above extracted rules only include 6 attributes: a_3, a_4, a_5, a_6, a_7 and a_8. Now we will see whether those 6 attributes can disambiguate the two senses of *must* with accuracy as high as the one yielded by the prior 17 attributes. In Table 5, the other 11 attributes are deleted and a new formal context with the 6 remained attributes is produced (see Table 5).

Table 5. Formal context of *must* with optimized features



No.	a_3	a_4	a_5	a_6	a_7	a_8	No.	a_3	a_4	a_5	a_6	a_7	a_8
o1(1)		1	1		1		o101(2)	1			1		1
o2(1)	1		1		1		o102(2)	1			1		1
o3(1)		1		1	1		o103(2)	1			1		1
o4(1)		1		1	1		o104(2)	1			1		1
o5(1)	1		1		1		o105(2)	1			1		1
…	…	…	…	…	…	…	…	…	…	…	…	…	…
o96(1)		1	1		1		o196(2)	1			1		1
o97(1)		1	1		1		o197(2)	1			1		1
o98(1)		1	1		1		o198(2)	1			1		1
o99(1)		1	1		1		o199(2)	1			1		1
o100(1)		1	1		1		o200(2)	1			1		1

This formal context is clarified and converted into SPOAD as well, generating the new WSD model with only 6 features, as shown in Figure 4, through which the sense classification and determination of *must* can be realized.

4.3. **Testing of the new WSD model.** To verify the effectiveness of the new model built by optimized features, the original formal context of the testing data set is also transformed into a new one by reserving the 6 features included in the extracted rules and removing the other 11 features. Then the new formal context of the testing data set is compared with the model and the accuracy of the new model is calculated as 97.5%, which is higher than the model generated with all the 17 candidate features.

4.4. **Result discussion.** It can be seen that in the second model, the features being used are reduced from 17 to 6 (as nearly as 200%), but the accuracy rate is 3 percent higher. We may come to the conclusion that the feature set a_3, a_4, a_5, a_6, a_7 and a_8 are the optimized feature set for the WSD of the two-sense *must*, and the other features are redundant features, in other words, they do contribute to the WSD of *must*, but when the optimized feature set exists, they are of little significance.

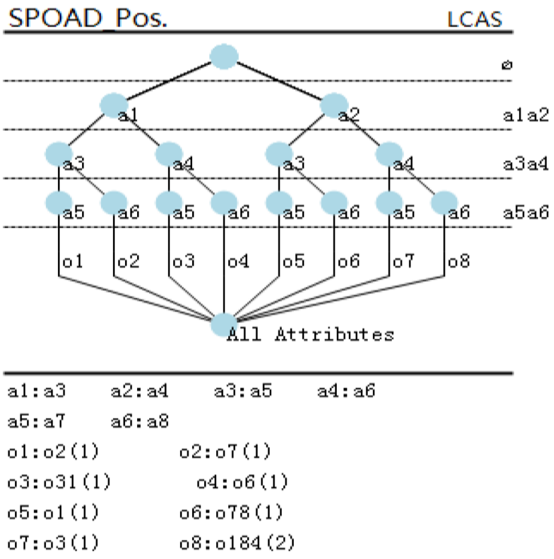


FIGURE 4. SPOAD of *must* with optimized features

5. **Conclusion.** This paper proposes a novel method to optimize the feature selection of English modal verb *must* in its WSD: constructing a structural partialordered attribute diagram with all candidate features and extracting rules from it, which generates optimized feature set. The follow-up experiment proves that the WSD model with optimized feature set is more effective and the accuracy rate is higher in terms of the two-sense WSD of *must*, i.e., it outperforms the original model with 17 features. The paper provides a new perspective for feature selection in binary classification in machine learning and natural language processing and might be applied to the research of other binary classifiers.

Acknowledgment. This work is supported by the Social Sciences Foundation of Hebei Province under Grant No. HB14YY005 and No. HB15YY002, the National Social Science Foundation under Grant No. 12BYY121, and the Humanities and Social Sciences Foundation of the Ministry of Education of China under Grant No. 14YJC740038. It is also partially supported by the National Natural Science Foundation of China under Grant No. 61273019. The authors gratefully acknowledge the supports.

REFERENCES

[1] G. Chandrashekar and F. Sahin, A survey on feature selection method, *Computers and Electrical Engineering*, vol.40, no.1, pp.16-28, 2014.

[2] R. Kohavi and G. John, Wrappers for feature subset selection, *Artificial Intelligence*, vol.97, nos.1-2, pp.273-324, 1997.

[3] I. Guyon and A. Elisseeff, An introduction to variable and feature selection, *Journal of Machine Learning Research*, vol.3, no.3, pp.1157-1182, 2003.

[4] L. Zou, Y. Pan, W. Zhu, G. Zhou and Y. Li, The object-oriented change detection based on neighborhood correlation images and minimum-redundancy-maximum-relevance feature selection, *Journal of Image and Graphics*, vol.19, no.1, pp.158-166, 2014.

[5] J. Xu, Y. Zhou, L. Chen and B. Xu, An unsupervised feature selection approach based on mutual information, *Journal of Computer Research and Development*, vol.49, no.2, pp.372-382, 2012.

[6] W. Cheng and X. Tang, A text feature selection method using the improved mutual information and information entropy, *Journal of Nanjing University of Posts and Telecommunications*, vol.33, no.5, pp.63-68, 2013.

[7] P. Pudil, J. Novovicova and J. Kittler, Floating search methods in feature selection, *Pattern Recognition Letters*, vol.15, no.11, pp.1119-1125, 1994.

[8] J. Reunanen, Overfitting in making comparisons between variable selection methods, *Journal of Machine Learning Research*, vol.3, no.3, pp.1371-1382, 2003.

[9] D. Dai and D. Mu, Heuristic genetic algorithm for feature selection in incomplete information systems, *Acta Electronica Sinica*, vol.41, no.3, pp.451-455, 2013.

[10] L. Chuang, H. Chang, C. Tu and C. Yang, Improved binary PSO for feature selection using gene expression data, *Computational Biology and Chemistry*, vol.32, no.1, pp.29-38, 2008.

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

[12] 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.

[13] 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.

[14] M. R. Perkins, *The Expression of Modality in English*, The Polytechnic of Wales, 1980.

[15] J. Coates, *The Semantics of the Modal Auxiliaries*, Routledge Press, London, 1983.

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

[17] 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.