Applied Soft Computing 27 (2015) 411–419

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

A new approach of rules extraction for word sense disambiguation by features of attributes

Jianping Yu[a,*], Chen Li[a], Wenxue Hong[b], Shaoxiong Li[b], Deming Mei[c]

[a] College of Foreign Studies, Yanshan University, No. 438 Hebei Street, Qinhuangdao 066004, Hebei, PR China
[b] Institute of Electrical Engineering, Yanshan University, No. 438 Hebei Street, Qinhuangdao 066004, Hebei, PR China
[c] College of International Programs, Shanghai International Studies University, No. 410 Dong Ti Yu Hui Road, Shanghai 200083, PR China

ARTICLE INFO

Article history:
Received 25 July 2013
Received in revised form 5 June 2014
Accepted 28 October 2014
Available online 5 November 2014

Keywords:
Rule extraction
English preposition
Feature of attribute
Word sense disambiguation

ABSTRACT

Classification is an important issue in data mining and knowledge discovery. It is a significant issue to develop effective and easy approach of rule extraction for classification. A new approach of rule extraction by features of attributes is proposed in this article for word sense disambiguation (WSD). English preposition *on* is taken as a target word of WSD, a data set of 600 samples is randomly selected from a 350,000 words corpus. Semantic and syntactic features are extracted from the context, and the corresponding formal context is generated. The rules for WSD of English preposition *on* are extracted based on the theoretical descriptions and calculation of the simple class exclusive attributes and composite class exclusive attributes. The extracted rules are used in the WSD of English preposition *on*, and the accuracy reaches 93.2%. The results of the comparative analysis show that the proposed feature of attribute approach is simpler, more effective and easier to use than the existing well-formed structural partial ordered attribute diagram approach.

1. Introduction

Rule extraction is an important issue in natural language processing. It is a process of deriving a symbolic description of a model for classification. It simulates the behavior of the model in a concise and comprehensible form. Rule extraction gives insight into the logic behind the model. Many researchers have studied rule extraction from different perspectives. In the aspect of rule extraction from different models, Setiono et al. [1] proposed an approach for rule extraction from minimal neural networks for credit card screening. Ozbakir et al. [2] proposed an approach for rule extraction from artificial neural networks to discover reasons of quality defects in fabric production. Chorowski and Zurada [3] presented an eclectic approach for rule extraction from neural network as decision diagrams. Zhu and Hu [4] proposed a rule extraction technique by support vector machines through analyzing the distribution of samples. Chaves et al. [5] proposed a new method for fuzzy rule extraction from trained support vector machines for classification of multi-class problems. Tang et al. [6] presented a method of extracting classification rules from concept lattice. Li et al. [7] extracted rules for word sense disambiguation (WSD) of English modal verbs from a structural partial ordered attribute diagram. Asaduzzaman et al. [8] reported a method of finding out interesting rules from heterogeneous internet search histories.

In the aspects of the algorithms for rule extraction, Liu et al. [9] proposed an algorithm to simultaneously extract rules and select features for better interpretation of the predictive model. Zhao and Sun [10] proposed an approach to rough set rule extraction from a decision system using conditional information entropy. He et al. [11] proposed a guidance rule extraction algorithm for getting the attribute information along the quickest direction and achieving the intelligent information analysis. Sun [12] developed an algorithm framework for rule extraction with different levels of knowledge granular from decision system in order to delete redundant features from decision system and highlight the most efficient features to construct classifiers. Ahmed and Carson-Berndsen [13] presented a method for automatic rule extraction for modeling pronunciation variation in order to model pronunciation variation in phoneme based continuous speech recognition at language model level. Sarkar et al. [14] introduced a genetic algorithm-based rule extraction system to improve prediction accuracy over any classification problem irrespective to domain, size, dimensionality and class distribution. They [15–17] also proposed a hybrid approach to design efficient learning classifiers and an accuracy-based learning system to extract efficient rule set for the implement of a multi-category classification, and select informative rules by using

* Corresponding author. Tel.: +86 15243477660.
E-mail address: yjp@ysu.edu.cn (J. Yu).

http://dx.doi.org/10.1016/j.asoc.2014.10.037
1568-4946/© 2014 Elsevier B.V. All rights reserved.

Nomenclatures

WSD	word sense disambiguation
TES*on*	"*on*" with the sense of "time"
NPS*on*	"*on*" with the sense of neither "time" nor "place"
PS*on*	"*on*" with the sense of "place"
n	noun
K	a formal context
G	a set of objects
M	a set of attributes
I	a set of relation between objects and attributes
g	an object
m	an attribute
A	extent of a concept
B	intent of a concept
M_s	a subset of M
G_s	a subset of G
Ø	empty set
G_1	a set of a class of objects
G_p	any set of objects other than G_1
D_p	decision attribute set
G_i	object sets corresponding to the decision attributes
M_c	non-decision attribute set of a class
M_{cj}	non-decision attribute set corresponding to G_i
g_j	the jth object in G_1
$P(M_{ci})$	power set of M_{ci}
M_{cik}	the kth attribute set of $P(M_{ci})$
G_{cik}	object set corresponding to m_{cik}
G_{c1i}	a subset of G_1
MI	mutual information
$P(w_1, w_2)$	probability of co-occurrence of w_1 and w_2
$P(w_1)$	probability of w_1
C_i	the ith concept
mi	the ith attribute
gj	the jth object
SPOAD	structural partial ordered attribute diagram

parallel genetic algorithm. Rodriguez et al. [18] presented an efficient distributed genetic algorithm for classification rule extraction in data mining. Wang et al. [19] introduced a method for rule extraction based on granular computing in order for default diagnosis of a helicopter transmission system. Koklu et al. [20] presented a new method of rule extraction from medical related datasets using artificial immune system algorithm. Huang et al. [21] proposed a method based on clustering artificial fish-swarm algorithm and rough set theory to extract decision rules. Costro et al. [22] described a rule extraction algorithm based on fuzzy logic, named linguistic rules in fuzzy inductive reasoning, to derive linguistic rules from a fuzzy inductive reasoning model. Chen et al. [23] presented an integrated mechanism for simultaneous extraction of fuzzy rules and selection of useful features in order to solve the classification problem. Cheng [24,25] studied the approached for rule extraction in fuzzy information systems based on rough set theory.

The previous studies in rule extraction have solved many practical problems and made a great progress in natural language processing. However, most of them have focused on the rule extractions for solving problems in engineering, business, medical diagnosis, and fuzzy information system etc. Up to now, few of them are related to WSD and no studies on the rule extraction by features of attributes have been found. In addition, there is a clear need in natural language processing to develop approaches which can extract effective and high quality rules for classification with less effort. Therefore, a new approach of rule extraction by features of attributes is proposed in this article for WSD of English preposition, with *on* as a target word, in order to simplify the process of rule extraction and improve the qualities of the extracted rules and the accuracy of WSD. The proposed approach may be applied to the WSD of other English prepositions, and it can also be used in different fields, such as pattern recognition, knowledge discovery, data mining, default diagnosis, decision support system and intelligent robot. The result of the study may provide references for natural language processing and understanding, semantic studies of prepositions and WSD of other part of speech.

The rest of the article includes the following contents. Section 2 presents the senses of *on* occurred in the corpus and the granularity of the senses of *on* in this study. Section 3 gives the theoretical descriptions of formal context and features of some attributes. Section 4 gives the procedure of calculating simple class exclusive attribute and composite class exclusive attributes. Section 5 explains the process of generation of the formal context of English preposition *on*. Section 6 exhibits the process of rule extraction for WSD of *on*. Section 7 makes a comparison between two approaches of rule extraction; the feature of attribute approach and the structural partial ordered attribute diagram approach. Finally, Section 8 comes to the conclusions of the study.

2. Granularity of the senses of English preposition *on*

English preposition *on* is one of the most frequently used simple prepositions in the natural language. It may have about 20 senses, and it may mean differently in different contexts. For instance [26],

(1) My mobile phone is ***on*** the table (*on*-place)
(2) The meeting will be ***on*** Tuesday (*on*-time)
(3) He made a lot of money ***on*** the deal (*on*-cause)
(4) He walked ***on*** tiptoe (*on*-manner)
(5) He had thrown me down ***on*** one hundred pitchforks (*on*-direction)
(6) He would lead a violent assault ***on*** the jail (*on*-objective)
(7) His role ***on*** base was hardly formal (*on*-subordination)

The ambiguity of *on* may cause trouble in natural language understanding and processing. In addition, the senses of prepositions are complex; therefore, the disambiguation of prepositions is a tough and inevitable issue. In this study, a corpus is constructed in order to extract rules for WSD of English preposition *on*. The corpus is composed of 350,000 words including written and spoken materials, such as research articles, laws, movie subtitles, news, speeches, literatures, and interviews, and about 50,000 words for each genre are included in the corpus. The senses of *on* in the corpus are tagged and the occurrence of *on* with each sense is counted by Wordsmith Concordance Tool, the statistic result is shown in Table 1. In this study, the granularity of the senses of *on* is set to be 3 in order to simplify the WSD and the rule extraction of it: (1) *on*-time, *on* with the sense of "time" (TES*on*); (2) *on*-others, *on* with sense of others—neither "time" nor "place (NPS*on*) and (3) *on*-place, *on* with the sense of "place" (PS*on*).