International Journal of Database Theory and Application
Vol.9, No.6 (2016), pp.93-102
http://dx.doi.org/10.14257/ijdta.2016.9.6.09

Semantic Role Labeling Based Event Argument Identification

Wen Zhou, Yajun Zhang, Xiaoying Su, Yao Li and Zongtian Liu

School of Computer Engineering and Science, Shanghai University
zhouwen@shu.edu.cn, zyj1985email@163.com, ying28227@163.com, 317286067@qq.com, ztliu@shu.edu.cn

Abstract

Event extraction is one of the most challenging tasks of information extraction from text. This paper studies one of the stages of Chinese event extraction, namely, event argument identification. A new method we call Semantic Role Labeling Based Event Argument Identification, based on the state-of-the-art methods of event extraction and event argument identification, is proposed. First, the 5W1H (who, what, whom, when, where, how) information is extracted from the text using semantic role labeling; thereafter, the 5W1H information is mapped to each argument of the event by heuristic rules. The method is used to identify the event arguments on two test sets of acquisition and transfer data, and contrasted with the methods of SRL and SRL combining heuristic rules. We find the best F1 measures for each to be 76.04% and 79.19% respectively.

Keywords*: Event argument identification, SVM, Semantic Role Labeling, 5W1H*

1. Introduction

Event extraction is an important topic of research in the field of text information extraction. Its aim is to display unstructured text, including event information, as structured text. Event extraction technology is widely used in automatic summarization [1-2], automatic answering of questions [3-4], and information retrieval [5-6]. In 2005, Automatic Content Extraction (ACE) added event extraction as part of the tasks. ACE defines an event as being composed of an event trigger and event arguments, and proposes that the event extraction task consists of event-type recognition and event argument recognition. The latter contains details including the time of occurrence of the event, place and participants. Figure 1 shows the annotation instance for event argument "acquisition" according to the ACE Chinese event annotation guide. In this instance, "acquisition" is the event trigger, the event type is "Business," and the sub-type is "Merge-Org." The four elements of the event are " company A", "Company B","2011" and "20% equity," which correspond to the three argument labels "Org", "Time" and "Object" of the event class.

2. Related Work

Semantic role labeling (SRL) [7] is now considered an integral step in semantic analysis of natural language sentences and especially in information extraction. The core idea of SRL is endowing these sentences with components-fixed semantic role information according to semantic relationships between the predicate and sentence components of different phrases in the sentences. SRL simply labels components related to the sentence predicate and treats these components as the predicate's

ISSN: 2005-4270 IJDTA

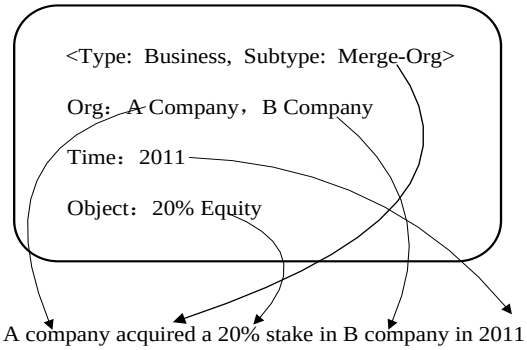

Figure 1. The Basic Elements of Acquisition Event

Parameters endowed with certain semantic meaning. Semantic role labeling integrates the core technologies in natural language processing such as word segmentation, part-of-speech tagging, syntactic analysis. Therefore, research on SRL for text lays a solid foundation for text structuring and intelligent information processing based on natural language. As a core technology in the field of natural language processing, SRL plays an important role in several applications. In particular, SLR provides a powerful means of answer the basic 5W1H (who, what, when, where, why and how) questions.

Prevalent event argument recognition methods can be divided into two types: methods based on models, and those based on machine learning. Model-based methods are primarily manually designed custom event models that use a variety of pattern matching algorithms to match the text with the custom model. For example, [8] proposes a football event information extraction system, and [9] develops a meteorology event extraction system based on ontology. Methods based on machine learning focus on the construction of a classifier, discovery of features and their combination, and selection. These methods approach event argument recognition as a classification problem and select appropriate features for their classification. In 2002, Chieu and Ng [10] first introduced a maximum entropy classifier for event extraction and used it to recognize event arguments. To address the fact that existing event extraction methods by filling events don't consider context, Huang *et al.* [11] proposed a bottom-up approach for event extraction that initially identifies candidate role fillers independently and then uses that information as well as discourse properties to model textual cohesion. The novel component of the architecture is a sequentially structured sentence classifier that identifies event-related story contexts. The sentence classifier uses lexical associations and discourse relations across sentences, as well as domain-specific distributions of candidate role fillers within and across sentences. This approach yields state-of-the-art performance on the MUC-4 dataset, achieving substantially higher precision than previous systems.

Dommati *et al*. [12] have focused on feature extraction, noise reduction in data and classification of network bugs using a probabilistic Naïve Bayes approach. Different event models like the Bernoulli or Multinomial models are applied on the extracted features. When new unseen bugs are given as input to the algorithms, the performance comparison of different algorithms is done on the basis of precision and recall parameters. Amid *et al.* [14] proposed a new approach to classify and rank multimedia events based purely on audio content using video data from the TRECVID-2013 multimedia event detection (MED) challenge [13]. Jiangfeng Fu *et al.* [14] proposed an event extraction method based on weight value feature of event arguments in event ontology.

In [15], Wang *et al.* propose a verb-driven method to extract event semantic information (5W1H) from Chinese texts and proved the reliability and feasibility of this method by using experimental data. Yankova and Boytcheva [8] propose an event extraction method that integrates machine learning, statistical methods and SRL based on the PropBank corpus. This method got satisfactory results for English language texts.

McCracken *et al*. [16] extract multiple arguments using a model matching method based on a syntax tree. Seokhwan *et al.* [17] use a method to combine lexical semantic and semantic roles in order to recognize and classify events automatically. However, due to a lack of test sets, further experiments are needed to verify the validity of this method.

In light of the shortcomings of the above methods, inspired by [15-16], a new event argument recognition method based on SRL is proposed in this paper. Experiments on two sets of "acquisition" and "transfer" events of an enterprise alliance are used to verify the performance of the proposed method.

3. SRL-Based Event Argument Recognition

Our proposed event argument recognition method based on SRL includes two modules: an SRL module and a heuristic rule set module. The method first annotates a given text using SRL and then extracts 5W1H information. Following this, it uses heuristic rules to map the 5W1H information to each argument of "acquisition" and "transfer" of the enterprise alliance events. Figure 2 shows a frame diagram of the event argument recognition system proposed in this paper. Compared to traditional event argument recognition methods, our proposed method does not require an extensive manual annotation corpus and thus simplifies event argument recognition. Furthermore, our method does not require a predefined event argument module for different events and, thus, a large amount of prior knowledge. Therefore, it has a considerable advantage over conventional event recognition methods in specialized fields where knowledge and resources may be scarce.

3. 1. SRL Module

The SRL module is divided into three parts: pre-processing, semantic role recognition and 5W1H mapping.

Preprocessing involves event type recognition tasks on a corpus in order to identify the verb predicates. As our key task in this paper is event argument recognition, we will not describe the event type recognition process. We use the HKUST Chinese Semantic Parser to annotate the pre-processed corpus, which is developed based on Sameer Pradhan's Automatic Statistical Semantic Role Tagger (ASSERT) tool for English text [19]. It is the most advanced tools of its kind and can show Chinese verb predicates and different kinds of phrases in PropBank tree form. We will introduce PropBank tree, Chinese PropBank tree and the SRL event argument recognition method in details in the following sections of this paper.

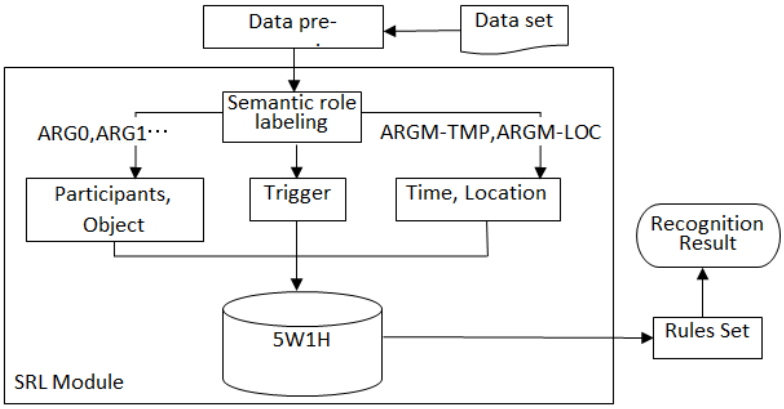

Figure 2. Event Argument Recognition System Frame Diagram

A. PropBank. The PropBank tree is currently well known as an English SRL corpus resource. It is a shallow semantic information set which the University of Pennsylvania annotated based on syntax analysis of the Penn TreeBank. The PropBank includes more than 50 semantic roles. Table 1 shows some labels and meanings of two kinds of semantic roles: semantic roles of predicate and connectivity semantic roles.

Table 1. Semantic Role Labeling Type for Propbank

The semantic role of Predicate	
ARG0	Usually express executant
ARG1	Usually express recipient
ARG2;3;4;5	Different meanings according to the different predicate
Connectivity semantic role	
LOC	Location
TMP	Time
NEG	Negative label
ADV	General purpose
CAU	Cause
EXT	Extent
DIR	Event direction
MOD	Modal verb
MNR	Mode, habit
PNC	Purpose
DIS	Event discuss

PropBank only annotates verbs (the target verb), and the same semantic roles may contain different semantic information because of different target verbs. Figure 3 gives an example of annotating one sentence, where [She] expresses the executant, [bought] expresses the target verb, [the silk] expresses the recipient, and [in China] expresses location.

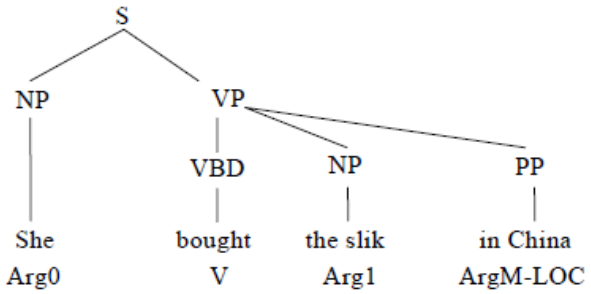

Figure 3. Propbank Tree Example

B. Chinese PropBank. Currently, in Chinese, corpus are annotated according to shallow semantic analysis, are few in number and the corpus primarily used is the Chinese PropBank. The source of the Chinese PropBank corpus is newspapers. It is composed of syntax trees generated following word segmentation and part-of-speech tagging. The corpus annotates the semantic information of all basic syntax trees, adds predicate-argument relations to them and then forms a shallow semantic parsing text corpus. In this paper, we use the structured syntax tree as a reference and manually annotate the raw corpus of the enterprise alliance field.

C. SRL. We manually collected all events containing "acquisition" and "transfer" from the Wind enterprise alliance corpus, re-annotated them according to the specifications of

the Chinese PropBank corpus and generated test sets and training sets compatible with the HKUST Chinese Semantic Parser[1] - namely, files in the "chtb100.fid" format. During the process of semantic role recognition, when we selected data sets ranging from "chtb100.fid" to "chtb931.fid" (Chinese PropBank1.0 format) as the training set and those from "chtb001.fid" to "chtb099.fid" as the test set for annotation experiments, the F-measure of the annotation results was a substantial 92.45%. In using the HKUST program package we replaced the original word segmentation tool developed by Stanford University with ICTCLAS, since the latter is the best available tool for word segmentation and part-of-speech (POS) tagging in Chinese. After lots of experiments, we discovered that it can improve role-labeling performance. With regard to feature selection, Pradhan *et al.* [20] summarized commonly used features, such as the predicate, path, phrase type, grammatical voice, position, verb sub-classes of the framework and the core word. These basic features reflect the semantic role labeling information of the object from different perspectives. In addition to these basic features, we add several extended features which can reflect partial information of units to be annotated. The feature set selection of HKUST Chinese Semantic Parser is as in Table 2.

D. Mapping 5W1H. The 5W1H mapping results from the SRL responses to each of the 5W1H questions: who, what, whom, when, where, and how. A response to "who" expresses the event executant, the "what" expresses the event type, the "whom" expresses the event recipient, the "when" expresses the time of the occurrence of the event, the "where" expresses the location of the event and a response to the "how" expresses the process of the event. As event argument recognition has no relation to the processing of the event, it's not mapped in this module. Table 3 shows the mapping results.

Table 2. The Feature Set of HKUST Chinese Semantic Parser

	Feature value	Feature description
Basic features	Predicate	Predicate itself
	Phrase and its Type	The phrase and the type of phrase
	Parent Phrase Type	The phrase type of the father node
	Head Word and its POS	The first character and the types of phrases
	Position	Whether the phrase is before or after the predicate
	Path	The expansion path from the phrase to the predicate
	Sub-category	Rules extension of parent node of the target verb
	Predicate + Head Word	Combination of predicates and the first character
	Predicate + HW POS	Part of speech combination of predicates and the first character
	Predicate + Phrase Type	Combination of predicates and phrases type
	First Word in Constituent	The composition components of the first character involved
	Last Word in Constituent	The composition components of the last character involved
Extended features	Left Path	The left half part of the path
	Right Path	The right half part of the path
	Half Path	Left Path with the phrase types of parent node
	Path Trigrams	The path of the triples
	Verb Cluster	The verb type information
	Path Abbreviation	Abbreviation of each phrase in the path

Table 3. Mapping Rules of Semantic Role Labeling Mapped to 5W1H

[1] http://hlt030.cse.ust.hk/research/c-assert/.

SRL Results	5W1H	Description
ARG-TMP	When	The time of the occurrence of the event
ARG-LOC	Where	The location of the event
Trigger	What	Event type
ARG0	Who	Event executant
ARG1	Whom	Event recipient

3. 2. Heuristic Rule Set

After we obtained the mapping results of the SRL and the 5W1H, we can roughly distinguish the corresponding arguments of the event, such as event time, location, object, participant, recipient and person from the content of each element of the 5W1H. However, since expressions in Chinese are very rich and complex, there are many event arguments which aren't able to clearly express an event. The most obvious examples are ARG0, ARG1, and other annotations in the SRL results. The HKUST Chinese Semantic Parser acquiescently annotates the participants prior to the predicates as ARG0, following which the predicates are annotated as ARG1, ARG2, ARG3 and so on. However, the method can't determine the role of the participant after the predicate. we define a series of heuristic rules for the "purchase" and "transfer" events in the financial industry to improve the accuracy and the recall rates of event argument recognition. The "how" isn't mentioned in event arguments recognition so "how" and its rule set are not considered here. By analyzing the features of the events from the corpus, we define a series of rules and methods to process such arguments, which are shown as Table 4.

Table 4. Problem Description and Rule Set

5W1H	Problem description	rule
What	One sentences annotated multiple verbs	To compute the semantic similarity between the verb (W_1) in the sentence and the core events' trigger (W_2) of this event class (acquisition class and transfer class) through HowNet, the specific computing formula is as follows: $Sim(W_1,W_2)=\max_{i=1..n,j=1...m} Sim(S_{1i},S_{2j})$, in which, S_{1i} and S_{2j} represent all meanings of W_1 and W_2 respectively.
Where	ARG-LOC appears a series of redundancy	To match the place phrases obtained from the ICTCLAS with the place phrases obtained through manual statistics from the corpus
Who	multiple ARGs appear before the predicate	If ARG is before the predicate it is considered an executant
Whom	multiple ARGs appear after the predicate	If ARG is after the predicate it is considered a recipient
When	ARG-TMP is missing	To match the place phrases obtained from the ICTCLAS with the time phrases of Date type and the event phrases obtained through manual statistics from the corpus [15]. If the matching phrase is a part of a time phrase in the corpus, we consider it as the time argument. If not, it is defined as null.

4. Experimental Results and Analysis

We collected various event reports and statements for April 2010 from the mergers and acquisitions corpus maintained by Wind Enterprises as our experimental data and completed the text pre-processing and annotation in advance. We used a semi-automatic annotation tool developed by the Intelligent Semantic Laboratory of Shanghai University to complete the annotation work. We annotated the event class and the event arguments. The event arguments included time (T), location (L), the executant (E), human name (HN), value (V), and recipient (O). There are more than 500 texts that contain the "acquisition" and "transfer" event classes. In the experiment, we used the HKUST

Chinese Semantic Parser which operates in the Linux environment as the SRL tool and we used ICTCLAS as the word segmentation tool. We analyzed the events of two types in the corpus, "acquisition" and "transfer", with the highest frequency of occurrence and used 10-fold cross-validation. The information extraction evaluation criteria included precision (P), the recall rate (R) and the F1-measure (F1) in order to assess the performance of the entire system. The criteria are defined as follows:

$$\text{Precision rate (P)} = \frac{\text{The total number of the arguments with correct recognition}}{\text{The total number of the arguments with efficient performance}}$$

$$\text{Recall rate (R)} = \frac{\text{The total number of the arguments with correct recognition}}{\text{The total number of the arguments with standard efficiency}}$$

$$\text{F1- measure (F1)} = \frac{2PR}{P+R}$$

Table 5 shows all event argument statistics for the "acquisition" and "transfer" events. From the table, we can observe that the executant (E), value (V) and recipient (O) appeared most frequently, time (T) and location (L) appeared less frequently and human name (HN) appeared with the lowest frequency. This is due to the nature of the corpus, which mainly recorded "acquisition" and "transfer" events among companies and due to "acquisition"-type behavior involving few people but a lot of price information.

Table 5. The Statistical Information of each Element in Corpus

Event argument type	"acquisition" event	"transfer" event
Executant(E)	289	269
Price(V)	278	253
Recipient(O)	254	234
Time(T)	65	53
Place(L)	50	48
Human Name(HN)	5	6

Table 6. Event Identification Result for "Acquisition" and "Transfer"

	SRL combination with heuristic rules			SRL		
Evaluation criteria	P	R	F1	P	R	F1
Executant (E)	78.20%	80.20%	79.19%	75.30%	76.80%	76.04%
Recipient(O)	77.30%	79.40%	78.34%	74.50%	76.10%	75.29%
Price(V)	70.40%	76.80%	73.46%	70.10%	75.30%	72.61%
Location(L)	68.60%	68.80%	68.70%	66.40%	67.90%	67.14%
Time(T)	65.20%	67.90%	66.52%	64.80%	66.70%	65.74%
Human Name(HN)	60.70%	62.40%	61.54%	59.80%	59.60%	59.70%

Table 6 shows the event argument recognition results of SRL and our proposed method of SRL combined with a heuristic rule set. From Table 6 we know that the overall event argument recognition results of using SRL combined with a heuristic rule set are better than results of only using SRL. The argument recognition results of the executant and the recipient in the corpus are the highest; the F1 value of the executants reaches 79.19% and 76.04%, respectively while the F1 values of human Name are 61.54% and 59.70%, respectively. Analyzing our experimental results we can draw the following conclusion:

(1) SRL combined with heuristic rules for event argument recognition performs better than simple SRL. Since there are few Chinese PropBank corpuses, this leads either to the incorrect annotation of certain attributes or the appearance of redundant annotation in SRL processing. Classification results are also affected while at the same time the feature set of SRL selection is not sufficiently abundant. However, a further restriction of the heuristic rules can compensate for the shortage of insufficient information of corpus in the specific field.

(2) Among the results of our experiments on event argument recognition using the two methods, recognition results for the executant, the recipient, and the price are better than those for the other corpus arguments. This is due to our experiment using the mergers and acquisitions corpus of Wind Enterprises as experimental data. The corpus naturally contains the names of several enterprises as well as acquisition participants and prices whereas human names and names of locations appear less frequently, affecting the SRL machine learning process.

5. Conclusions

In this paper, we propose a method that combines SRL with heuristic rules to accurately recognize event arguments in Chinese language text. Compared to traditional event argument recognition methods based on machine learning, our method doesn't require the processing of a massive manual annotation corpus, reducing the workload of system implementation. Furthermore, since our method combines SRL with a heuristic rule set, recognition results are more accurate. However, because the current Chinese PropBank corpus is not comprehensive and domestic SRL is not mature, we can't evaluate performance of this method in detail. We can obtain primary experiment results. In further research, we will aim to conduct more work on the performance of SRL, including selecting more abundant and effective feature sets and improving the Chinese PropBank corpus.

Acknowledgements

This paper is supported by National Science Foundation of China (No.71203135).

References

[1] S. Min, B. Dong and L. Xu, "An Improved Method for the Feature Extraction of Chinese Text by Combining Rough Set Theory with Automatic Abstracting Technology", Communications in Computer & Information Science, **(2012)**, pp. 496-509.
[2] R. He, B. Qin, and T. Liu, "A Novel Approach to Update Summarization Using Evolutionary Manifold-Ranking and Spectral Clustering", Expert Systems with Applications An International Journal vol. 39, no. 3, **(2012)**, pp. 2375-2384.
[3] Y. S. Jane, "A support vector machine-based context-ranking model for question answering", Information Sciences, vol. 224, no. 2, **(2013)**, pp. 77-87.
[4] Y. G. Cao, F. Liu and P. Simpson, "Ask HERMES: An online question answering system for complex clinical questions", Journal of Biomedical Informatics, vol. 44, no. 2, **(2011)**, pp. 277-288.
[5] C. Claudio and G. Romano, "A Survey of Automatic Query Expansion in Information Retrieval", ACM Computing Surveys, vol. 44, no. 1, **(2012)**, pp. 159-170.
[6] G. Travis, G. Navarro and S. J. Puglisi, "New Algorithms on Wavelet Trees and Applications to Information Retrieval", Eprint Arxiv, vol. 426, no. 5, **(2011)**, pp. 25-41.
[7] M. Surdeanu, X Carreras, P. R. Comas and L. Marquez, "Combination Strategies for Semantic Role Labeling", Journal of Artificial Intelligence Research, vol. 29, no. 1, **(2011)**, pp. 105-151.
[8] Y. Milena and S. Boytcheva, "Focusing on Scenario Recognition in Information Extraction", Conference on European Chapter of the Association for Computational Linguistics, **(2003)**.
[9] L. C. Shing, Y. J. Chen and Z. W. Jian, "Ontology-based fuzzy event extraction agent for Chinese e-news summarization", Expert Systems with Applications, vol. 25, no. 3, **(2003)**, pp. 431-447.
[10] H. L. Chieu and H. T. Ng, "A maximum entropy approach to information extraction from semi-structured and free text", Eighteenth national conference on Artificial intelligence American Association for Artificial Intelligence, **(2002)**, pp. 786-791.

[11] H. Ruihong and E. Riloff, "Modeling Textual Cohesion for Event Extraction", Twenty-Sixth AAAI Conference on Artificial Intelligence, **(2012)**.
[12] S. J. Dommati, R. Agrawal, M. R. G. Ram, S. S. Kamath, S. J. Dommati, and R. Agrawal, "Bug Classification: Feature Extraction and Comparison of Event Model using Naive Bayes Approach", Eprint Arxiv, **(2013)**.
[13] E. Amid, A. Mesaros, K. J. Palomaki, J. Laaksonen and M. Kurimo, "Unsupervised feature extraction for multimedia event detection and ranking using audio content", 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, **(2014)**, pp. 5939-5943.
[14] J. Fu, Z. Liu, Z. M. Zhong and J. Shan, "Chinese Event Extraction Based on Feature Weighting", Information Technology Journal, vol. 9, no. 1, **(2010)**, pp. 184-187.
[15] W. Wang, D. Zhao, L. Zou, D. Wang and W. Zheng, "Extracting 5W1H Event Semantic Elements from Chinese Online News", Web-Age Information Management, **(2010)**, pp. 644-655.
[16] N. E. Mccracken, N. Ozgencil and S. Symonenko, "Combining Techniques for Event Extraction in Summary Reports", Proceedings of AAAI Workshop Event Extraction & Synthesis, **(2006)**.
[17] S. Kim, M. Jeong, and G. G. Lee, "A local tree alignment approach to relation extraction of multiple arguments", Information Processing & Management, vol. 47, no. 4, **(2011)**, pp. 593-605.
[18] L. Hector, E. Saquete, and B. N. Colorado, "Applying semantic knowledge to the automatic processing of temporal expressions and events in natural language", Information Processing & Management vol. 49, no. 1, **(2013)**, pp. 179-197.
[19] Martin, and J. Dan, "Shallow Semantic Parsing Using Support Vector Machines", In Proceedings of the 2004 Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-04, vol. 63, no. 2, **(2004)**, pp. 77-98.
[20] S. Pradhan, K. Hacioglu, V. Krugler, W. Ward, J. H. Martin, and D. Jurafsky, "Support Vector Learning for Semantic Argument Classification", Machine Learning, vol. 60, no. 1-3, **(2005)**, pp. 11-39.

Authors

Wen Zhou, Ph.D., Associate Professor at School of Computer Engineering and Science, Shanghai University. Her research interests include Data Mining, Natural Language Processing and Complex System.