International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.8, No.3 (2015), pp.177-186
http://dx.doi.org/10.14257/ijsip.2015.8.3.16

Chinese Discourse Segmentation Based on Punctuation Marks

Yancui Li [1,2], Hongyu Feng [2]and Wenhe Feng[3,2*]

1 Department of Computer Science and Technology, Soochow University, Suzhou, China, 215006
2 Henan Institute of Science and Technology, Xinxiang, China, 453003
3 Computer School of Wuhan University, Wuhan, China, 430072
Yancuili@gmail.com, 332934328@qq.com, wenhefeng@gmail.com

Abstract

This paper addresses Chinese discourse segmentation based on punctuation mark. Particularly, we propose various kinds of lexical, syntactic, position and punctuation features to train classifiers for Chinese discourse segmentation. Experimental results on CDTB (Chinese Discourse Treebank) show that our method based on punctuation mark is appropriate for Chinese discourse segmentation with 89.2% in accuracy.

Keywords: *Chinese Discourse Segmentation, Punctuation Marks, Lexical Features, Syntactic Features*

1. Introduction

The natural language units can be divided into words, phrases, sentences and discourses. In discourse parsing, discourse refers to the whole language units connecting together various entities and eventualities appearing in the text. It is well-known that interpretation of a text requires understanding its relation and hierarchy since discourse units rarely exist in isolation. Research in discourse parsing has been drawing more and more attention in recent years due to its importance in various NLP applications, such as summarization [1-2], question answering [3-4], and dialogue generation [5].

Discourse parsing includes many tasks such as discourse segmentation, discourse relation classification and discourse structure construction. The first stage of discourse parsing is discourse segmentation, which segments a given discourse to Elementary Discourse Units (EDU) automatically. We use Chinese Discourse Treebank [6] for discourse segmentation. We know that the discourse is segmented by punctuation from the definition of the EDU. Punctuation is an important mark in written language, the same punctuation tends to have different syntactic or discourse function.

Our discourse segmentation task is actually a punctuation disambiguation problem. There has been amount of researches on Chinese punctuation from the view of natural language processing. For example, Jin *et al.* [7] proposes a method for classifying commas in Chinese sentences by their context, and then segments a long sentence according to the classification results. And after sentence segmentation, the dependency parsing accuracy is improved by 9.6%. Li *et al.* [8] studies the usage and function of Chinese punctuations in syntactic parsing. The idea is to split a long sentence into segments, and then parses them individually and reconstructed the syntactic parser for the original sentence. Xue and Yang [9] describe a method for disambiguating Chinese commas that is central to Chinese sentence segmentation. Chinese sentence segmentation is viewed as the detection of loosely coordinated clauses separated by commas. Train and test on the data which is derived from the Chinese Treebank,

ISSN: 2005-4254 IJSIP

the accuracy of their model is close to 90% overall. Yang and Xue [10] proposes an approach to disambiguate the Chinese comma. Training and testing data are also automatically extracted from the Chinese Treebank based on several given syntactic patterns. They extract features from automatic parsers to train a classifier. From above we can see that sentence segmentation based on punctuation is a common method. Because lack of discourse corpus, related Chinese researches mainly focuses on the automatic extraction of syntactic pattern as training and testing data, there have no true discourse segmentation research especially.

There are 16 punctuation marks used in Chinese commonly, with point mark and label mark categories (GB/T15834) [11]. The role of point mark is the middle sentence and end sentence punctuation. The end sentence punctuations are period, question mark and exclamation mark, which represent the end of the sentence pause. The middle sentence marks are comma, semicolon and colon, which represent a variety of different nature pause within the sentence. The role of label mark is marked, which mainly marks the nature of the statement. There are 9 commonly used label marks, namely quotes, brackets, dashes, ellipsis, emphasis, connection number, interval number, name and the names of books. As a matter of fact, there are varieties of punctuations in the corpus, and the punctuations also have different effect. Inevitable for the EDU boundary, punctuations have great significance for sentence segmentation. There is certain relationship between Chinese written language and EDU boundary, that the period, question mark, exclamation mark and semicolons are boundary of EDU, while comma and colon are possible boundary of EDU. The frequency of punctuations in CTB6.0 from Li *et al*. [12] show that definite EDU boundary punctuation marks of period mark, question mark, exclamation mark and semicolon occupy 31.1%. While possible EDU boundary punctuation occupies 68.9%, with comma occupies 67.2%. So the key problem of discourse segmentation is to judge whether the punctuation is the boundary of EDU.

Punctuation is very important for discourse segmentation. Li *et al.* [12] analyses the relationship between the comma and EDU, and researches EDU segmentation using comma on annotation corpus. Experiments show that the definition of clause is reasonable and the identification of clause based on the comma is feasible. Inspired by their research, in this paper, we first introduce the Chinese discourse Treebank, especially the annotation of EDU based on punctuation. Then we introduce the experiment method, including the framework, the features and experiment setting. Finally we give the experiment results and conclude our work.

2. Chinese Discourse Treebank

For Chinese discourse, to our knowledge, there hasn't well-established corpus which is available for Chinese. According to RST-DT, PDTB, Chinese complex sentence[13] and sentence-group theory [14], We adopt a presentation format of connective dependency tree, in which leaves are EDUs and intermediate nodes are connectives, as annotation scheme for Chinese discourse tree bank (CDTB).For detail you can reference Li *et al.* [6]

Example (1): 1浦东开发开放是一项振兴上海，建设现代化经济、贸易、金融中心的跨世纪工程，||2因此大量出现的是以前不曾遇到过的新情况、新问题。| 3 对此，浦东不是简单的采取“干一段时，等积累了经验以后再制定法规条例”的做法，|||4而是借鉴发达国家和深圳等特区的经验教训，|||| 5聘请国内外有关专家学者，||||6积极、及时地制定和推出法规性文件，|||7使这些经济活动一出现就被纳入法制轨道。||8去年初浦东新区诞生的中国第一家医疗机构药品采购服务中心，正因为一开始就比较规范,|||9运转至今,||||| 10成交药品一亿多元，|||| 11没有发现一例回扣。

Pudong development and opening up is a cross-century project of promote Shanghai, building a modern economy, trade and financial canter. ||2 Therefore, there are a large

number of new situations and new problems that have not previously been encountered. | 3 Pudong is not only simply adopting "does a period of time, wait accumulation of experience then develop laws and regulations" approach to this. ||| 4 But also learns lessons from developed countries and the Shenzhen Special Administrative Region. |||| 5 Employ relevant experts and scholars at home and abroad. |||| 6 Actively and timely formulate and launch the legal document. |||7 So that economic activities can be incorporated into the legal system when they appeared. || 8 China's first drug procurement service canter of medical institutions, born in the Pudong New Area at the beginning of the last year, just because relatively standard at beginning ,||| 9 operated up to now, |||| 10 deal drugs more than one hundred million Yuan,||| 11 have not been found a case of kickbacks.(chtb_0001)

There are three sentences in example (1), "|" indicates first layer,"||" indicates second layer,"|||" third layer and so on. Arabic numerals indicate EDUs. We bold the connective words for emphasis. The discourse parser tree of example (1) is shown in Figure 1:

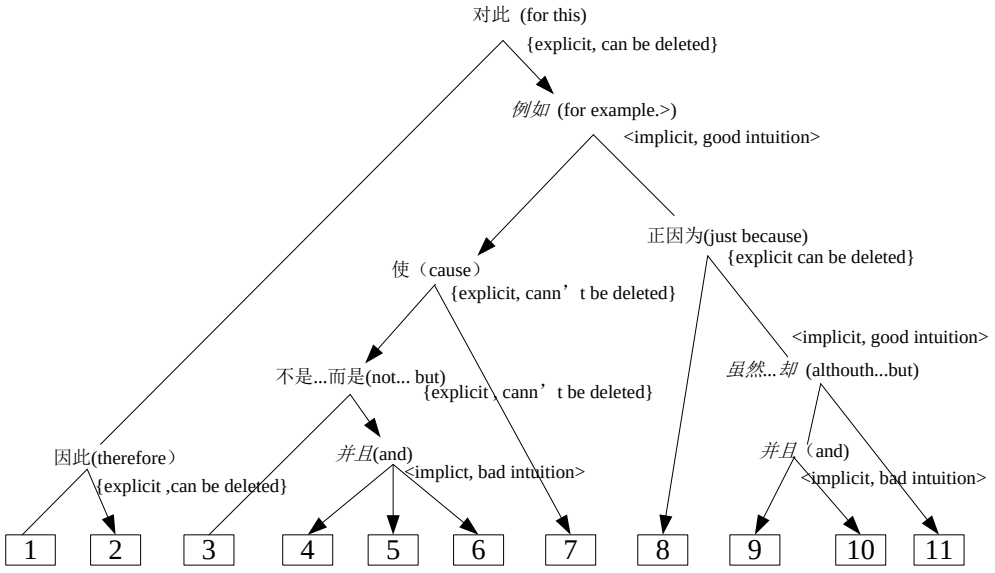

Figure 1. An Instance of Example1's Discourse Dependence Tree

Figure 1 shows that, the nodes of the discourse tree are constructed by EDUs and connectives. Leaf nodes in Figure 1 (Number marks such as 1, 2 etc.) indicate the EDUs, while non-terminal nodes are connectives. The combination of different EDUs can be considered as EDUs in a higher level and then the new discourse unit can be combined to higher units again. Finally the discourse can be expressed as a hybrid tree of EDUs and connectives. In our discourse parser trees, connectives can not only represent the discourse relation, but also represent the discourse hierarchical structure in the tree. The arrows in the tree point to the main EDU or main discourse unit. The discourse parser tree is very like dependency parser tree, and this is why we call it "Connective-driven dependency tree".

In Chinese Discourse Treebank (CDTB), we use our representation format to build discourse tree for each paragraph. We will introduce the annotation method of EDU and the scale of corpus in detail as follow.

EDU in CDTB is usually clause, including traditional simple sentence and clause in complex sentence. EDU contains at least one prediction, expresses at least one proposition, and must be segmented by some punctuation, usually commas, semicolons, and periods. The example (1) is divided by this definition and number mark indicates EDU.

In CDTB, we annotate whether the punctuation (period, question mark, exclamation mark, comma, semicolon, or colon) is the boundary of EDU, the layer of it and the attributes of it. Take the second comma in example (1) for example, the comma is the EDU boundary, the layer is 2, it is explicit relation, the connective is “因此 (so)”, the nuclear is right EDU of number 2.

Currently, the CDTB corpus consists of 500 newswire articles from Chinese Treebank, which are further divided into 2342 paragraphs with a CDT representation for one paragraph. For EDUs, CDTB contains 10650 EDUs with an average of 4.5 EDUs per tree. On average, there are 2 EDUs per sentence and 22 Chinese characters per EDU. The agreement of discourse segmentation for our corpus is 91.7%, and the Kappa value [15] is 0.91.

3. Method

The overall accuracy of discourse parsing depends on the segmentation result. If the text is wrongly segmented during the first stage, it becomes unreliable to build a consistent discourse tree for the text. Therefore, the discourse segmentation task is very important for discourse parsing.

Our segmenter implements a binary classifier to decide for each punctuation in the text, whether it is the boundary of an EDU or not. From section 2 we can know, punctuation which is the possible boundary of EDU is annotated whether it is an EDU boundary in our corpus. In discourse segmentation, the problem is to assign the punctuation of input text an observation category {+1,-1}, where “+1” indicates that punctuation is a boundary, and “-1” indicates that punctuation is not a boundary. For Example (1), the first comma is “-1” category, whereas the second comma belongs to category “+1”. Hence, we can model the discourse segmentation problem as binary classification and train a classifier. Then we use the classifier to obtain a list of EDUs from input text.

3.1. Framework

This paper mainly researches discourse segmentation based on punctuation using the supervised method. The framework of our discourse segmentation pipeline is shown in Figure 2.

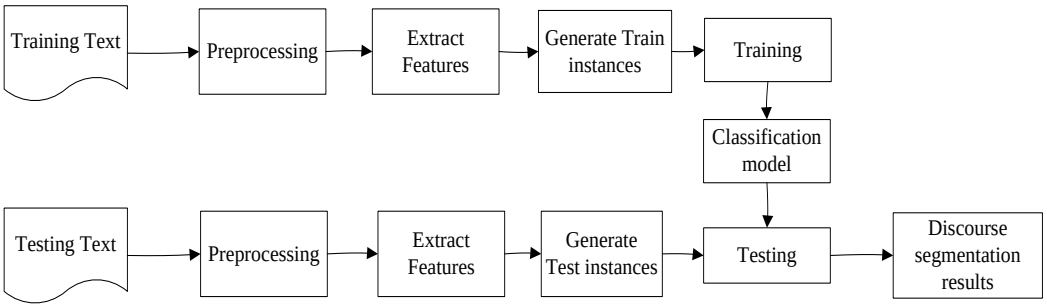

Figure 2. The Framework of Discourse Segmentation

Training texts are firstly under preprocessing, which include obtaining the punctuation boundary mark, word segmentation and part-of-speech tagging, syntactic analysis. Then the lexical, syntactic features and position features are extracted to get training instances, and then training obtains the classification model. After preprocessing, extract features from testing texts to generate test instances. Then the classification model is used to classify the

testing instances and predicate whether the punctuation is EDU boundary. Finally the discourse segmentation results are outputted.

3.2. Features

For discourse segmentation task we apply features described in Xue and Yang[9] for sentence segmentation as punctuation classification, and we also give a set of syntactic and lexical features as follow, each punctuation has two Spani and Spanj labels to indicate the left and right span separately:

Cue phrase

It is proved that cue phrase is very important for English discourse parsing, and many systems adopt it. But there is no collected connective table for Chinese discourse parsing to our knowledge. So we collect a cue phrase table containing Chinese connective words based on the principle of complex sentence research result. For each cue phrase in list, we determine whether it appears in Spani or Spanj. We also determine whether its appearance is in the beginning, the end or the middle of that span.

Lexical features

1 The last three words and their part of speeches of Spani; the first three words and their part of speeches of Spanj

2 The combination of the first word and the last word of Spani or Spanj;the combination of the first word part of speech and the last word part of speech of Spani or Spanj

3 Verbs and their part of speeches in Spani and Spanj. Here a word is verb if its part of speech is VV, VC, VE, or VA

4 Averbs in Spani and Spanj

5 Conjunctions in Spani and Spanj

6 Common words and their part of speeches between Spani and Spanj

Syntactic features

1 The phrase label of the Spani ; the phrase label of Spanj; the combination of the phrase label of Spani and Spanj

2 The combination of the punctuation's parent phrase label and the phrase label of Spani and Spanj

3 Whether the phrase label of Spani and Spanj are labeled as IP

4 Whether the punctuation is a child of the root node in the syntactic tree

5 The layer of the punctuation in the parser tree numbered from root

Position and punctuation

1 Whether the length of Spani is less than 5. Whether the length difference between Spani and Spanj is smaller than 7

2 The punctuation (exclude the end of sentence punctuation) of the sentence; the combination of punctuations in this sentence

3 The positions of the Spani and Spanj relative to paragraph boundaries (e.g., beginning, middle or end).

3.3. Experiment Setting

From CDTB corpus we can get every punctuation annotation of whether it is EDU boundary. There are 15485 punctuations in CDTB, with 10960 are EDU boundary and 4525 are not EDU boundary. EDU boundary accounts for 70.8% mainly because they include end sentence punctuation, such as period, semicolon, question mark, and exclamatory mark, and they definitely are EDU boundary. There are 9713 commas in CDTB, with 5436 are EDU boundary and 4277 are not EDU boundary. EDU boundary commas account for 56.0%. The

experiment mainly analyzes the performance of automatic segmentation on EDU. The experiment extracts the features described in section 3.2, uses the Decision Tree classifier, Maximum Entropy classifier and Naive Bayes classifier from mallet[16] respectively, and adopts 10 fold cross validation for discourse segmentation. Because period, semi-colon, question mark and exclamatory mark indicate the EDU boundary, so the experiment excludes these punctuations. There are 9949 mid-sentence punctuations (comma, colon, dash etc.), with 5486 are EDU boundary and 4463 are not EDU boundary.

4. Experiment Results and Analysis

4.1. Results of All Features

Using method described in section 3, this section gives the experiment results and analysis. We give the results of all features and individual features separately. The accuracy of inner-sentence punctuation segmentation and F-score for positive and negative instance respectively are shown in Table 1. In order to verify the effect of our features, we re-implement the work of [**Error! Bookmark not defined.**]. Because Xue *et al.* [**Error! Bookmark not defined.**] recognizes the comma which are function as period, while we recognize the punctuation as EDU boundary, the experiment results can't compare directly. We use the features of [**Error! Bookmark not defined.**] in our discourse segmentation task, and compare our features with their features both for our discourse segmentation experiment. We mark the punctuation which is EDU boundary as positive instance, the F1-score of it as F1(+), and the punctuation which is not EDU boundary as negative instance, the F1-scoure of it as F1(-).

Table 1. The Results of Discourse Segmentation based on Inner-sentence Punctuation

Classifier	Our features		Xue's features	
	Standard parse tree	Automatic parse tree	Standard parse tree	Automatic parse tree
	Acc. F1(+) F1(-)	Acc. F1(+) F1(-)	Acc. F1(+) F1(-)	Acc. F1(+) F1(-)
Maximum Entropy	**91.1 91.8 90.7**	**89.2 90.3 88.2**	**88.8 90.5 87.1**	**88.8 90.2 86.9**
Decision Tree	90.7 90.7 90.1	88.7 90.0 87.7	88.8 90.1 87.8	88.2 90.1 87.0
Naive Bayes	89.0 89.8 88.7	88.0 89.0 86.9	87.1 88.8 86.8	87.0 88.2 86.6

As shown in Table 1, standard parser means the parsers given in the CTB6.0 corpus, while automatic parser means the parsers produced by Berkeley parser. From table 1 we can see that the best experiment result accuracy is 91.1% for possible EDU boundary punctuation using standard parser, while using automatic parser for possible EDU boundary punctuation the accuracy is 89.2%. Comparing the results of our features with Xue's features, the accuracy of our features is 2.3% higher by using standard parser, this illustrates the features of ours are very effective. The performance of Maximum Entropy classifier is the best in these three classifiers, the accuracy is 91.1% when using standard parse tree and the accuracy is 89.2% when using automatic parse tree. The F-scores of that punctuation is EDU boundary (F1(+)) are 91.8% and 90.3% when using standard and automatic parse tree respectively. The F-score of punctuation not EDU boundary (F1 (-)) are 90.7% and 88.2% when using standard and automatic parse tree respectively. Comparing F1 (+) with F1 (-), we can see that the performance of punctuation not EDU boundary is better than that punctuation is EDU boundary. Comma is very important in discourse segmentation, and Table 2 shows the results of discourse segmentation based on comma.

Table 2. The Results of Discourse Segmentation based on Comma

Classifier	Standard parse tree			Automatic parse tree		
	Accuracy	F1(+)	F1(-)	Accuracy	F1(+)	F1(-)
Maximum Entropy	**91.5**	**92.4**	90.1	**88.4**	89.6	86.5
Decision Tree	91.2	92.3	90.8	88.3	89.6	86.2
Naive Bayes	87.8	89.4	85.7	84.5	86.7	81.2

As shown in Table 2, the performance of Maximum Entropy classifier is the best among the three classifiers. Using the standard parser from CTB6.0, the accuracy is 91.5%. Using the automatic parser produced by Berkeley parser, the accuracy is 88.4%. For the comma is EDU boundary, the F1 (+) is 92.4% by using standard parser and the F1(+) is 89.6% by using automatic parser. For the comma is not EDU boundary, the F1(-) is 90.1% by using standard parser and the F1(-) is 86.5% by using automatic parser. From Table 2 we can see that the performance of that comma is EDU boundary is better than that comma is not EDU boundary. The reasons are as follows: 1) Positive instances and negative instances are imbalance, with 56 percent viewing it positively and 44 percent negatively. 2) Because the punctuations not EDU boundary are complex, such as segmenting subject and predicate , verb and object of inner sentence, can't find useful features so that classification is very hard.

4.2. Results of Individual Features

We use many types of features for discourse segmentation, but there are some features important for discourse segmentation while some are not important. Table 3 gives the individual feature performance of our discourse segmenter.

Table 3. The Performance of Individual Feature

Features	accuracy	F1(+)	F1(-)
All	89.2	90.3	87.9
Cue phrase	72.4	82.1	21.1
Lexical features	82.9	86.3	71.8
Syntactic features	88.2	88.3	84.7
Position and punctuation	76.4	83.9	45.0

Table 3 shows that syntactic features contribute most in EDU recognition, followed by lexical features. The accuracy can reach 89.2%, the reason is that most of our EDU's labels are IP, VP, Coordinate IP and Coordinate VP in parser tree, and most of the NP, PP and LCP are not EDUs. Cue phrases we extract are the commonly used Chinese connectives, and the result shows that it is useful for determining the punctuation which is the boundary of discourse unit since the F1 (+) is 82.1%. Lexical feature is useful for either boundary or not boundary of the punctuation. Position and punctuation features are also useful. Especially for the feature length of Spani is less than five words are mainly not boundary.

4.3. Error Analysis

There are about 10% punctuation recognition errors, and we will analyze the reason of these errors as follows. There are two cases that negative instances are recognized as positive ones and positive instances are recognized as negative ones.

1）Negative recognized as positive

From the results of section 4.1, we can know that the negative punctuation recognition is lower in effect. The error situation mainly includes: punctuation is segmentation of subject and predicate; punctuation is segmentation of verb and object; punctuation is segmentation of adverbial. The examples are as follows:

Example (2):出口快速增长，(c1)成为推动经济增长的重要力量。

Export grew rapidly,(c1) which became important strength in promoting the economy to grow.（chtb_0097）

Example (3):确立了以资源换技术，(c2)以产权换资金，(c3)以市场换项目，(c4)以存量换增量的利用外资新思路。

It has established new thinking for utilizing foreign funds, such as exchanging resources for technology, (c2) exchanging property rights for capital, (c3) exchanging markets for projects and (c4) exchanging deposits for increments (chtb_0091)

Example (4):天津港保税区投入运行五年来，(c5)已建成了中国第一货物分拨中心，具备了口岸通关的功能，开通了天津港保税区经西安、兰州到新疆阿拉山口口岸的铁路专用线。

Since the Tianjin Port Bonded Area being put into operation five years ago,(c5) it has completed the construction of China's first goods distribution center, functions like a customs port, opened up the special use the railway line from the Tianjin Port Bonded Area passing Xi'an and Lanzhou to arrive at Xinjiang's Allah Mountain pass customs port. (chtb_0099)

Front of Comma c1 in Example (2) is subject of the whole sentence, so the comma is not EDU boundary, while syntactic analysis result of "出口快速增长(Export grew rapidly)" is IP, which usually represents a single sentence to make an error. The commas c2, c3, and c4 in Example (3) are between verb and object in sentence, but no feature can represent this information, so they are recognized as EDU boundary. The front span "天津港保税区投入运行五年来(Since being put into operation five years ago)" of c5 is the adverbial of the sentence in Example (4), c5 is the pause in adverbial, but "天津港保税区投入运行五年(the Tianjin Port Bonded Area being put into operation five years ago)" is an EDU, this makes the error.

2）Positive recognized as negative

Example (5):内地经济长期稳定地增长，(c6)香港经济将从充满活力的内地经济中获益。

The inland economy has been growing steadily in the long term (c6) and Hong Kong's economy will benefit from the vigorous inland economy. (chtb_0093)

The front of comma c6 in Example (5) is verb word "增长 (grow)", the after of comma c6 is Noun word "香港(Hong Kong)", this can produce the wrong recognition result.

5. Conclusion and Perspectives

In this paper, we aim to develop a Chinese discourse segmenter. By using the Chinese Discourse Treebank (CDTB) corpus, we present a discourse segmenter that segment text automatically based on punctuation mark. The discourse segmenter accuracy reaches 89.2% when using maximum entropy classifier and cue phrase, lexical and syntactic features. Experimental results show that our features are useful for discourse segmentation and the discourse segmentation based on punctuation is feasible.

Our future work will get more efficient feature to improve the discourse segmentation performance. Finally implement an end-to-end discourse parser containing discourse segmentation, relation classification and discourse tree building.

Acknowledgements

This research is supported by the National Natural Science Foundation of China No.61273320, by the Postdoctoral Science Foundation of China No.2013M540594,and by the Science and Technology Research Project of Henan Province Office of Education No.14A520080.

The contact author of this paper, according to the meaning given to this role by Wuhan University, is Wenhe Feng.

References

[1] D. Marcu, "The Theory and Practice of Discourse Parsing and Summarization", (2000), MIT Press.
[2] A. Louis, A. Joshi and A. Nenkova, "Discourse indicators for content selection in summarization. Proc. of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue", (**2010**), Portland, USA, pp. 147-156.
[3] S. Verberne, L. Boves, N. Oostdijk and P. A. Coppen, "Traitement Automatique des Langues", Discours et document: traitements automatiques, vol. 47, no. 2, (**2007**), pp. 21-41.
[4] R. Prasad and A. Joshi, "A Discourse-based Approach to Generating Why-Questions from Texts", Proc. of the Workshop on the Question Generation Shared Task and Evaluation Challenge, (**2008**), Arlington, VA.
[5] H. Hernault, P. Piwek, H. Prendinger and M. Ishizuka, "Generating dialogues for virtual agents using nested textual coherence relations", Proc. of the 8th international conference on Intelligent Virtual Agents, (**2008**), Tokyo, Japan, pp. 139-145.
[6] Y .C. Li, W. H. Feng, J. Sun and G. D. Zhou, "Building Chinese Discourse Corpus with Connective-driven Dependency Tree Structure", Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), (**2014**), Doha, Qatar, pp. 2105–2114.
[7] M. X. Jin, K. Mi-oung and K. Dongil, "Segmentation of Chinese Long Sentences Using Commas", Proc. of the SIGHANN Workshop on Chinese Language Processing. (**2004**), pp. 1-8.
[8] X. Li, C. Q. Zong and R. L. Hu, "A Hierarchical Parsing Approach with Punctuation Processing for Long Sentence Sentences", Proc. of the Second International Joint Conference on Natural Language Processing, (**2005**), pp. 17-24.
[9] N. W. Xue and Y. Q. Yang, "Chinese sentence segmentation as comma classification", Proc. of the 49th Annual Meeting of the Association for Computational Linguistics, (**2011**), Portland, Oregon, pp. 631-635.
[10] Y. Q. Yang and N. W. Xue, "Chinese Comma Disambiguation for Discourse Analysis". Proc. of the 50th Annual Meeting of the Association for Computational Linguistics, (**2012**), Jeju, Republic of Korea, pp. 786–794.
[11] China national standardization management committee, "GB/T15834-2011 Punctuation Usuage", (**2011**), Standards Press of China.
[12] Y. C. Li, W. H. Feng, G. D. Zhou and K. H. Zhu, "Research of Chinese Clause Identification Based on Comma", Acta Scientiarum Naturalium Universitatis Pekinensis, vol. 49, no.1, (**2013**), pp.7-14 (in Chinese with English abstract).
[13] F. Y. Xing, "Research of Chinese complex sentence", (**2003**), The Commercial Press.
[14] Z. Cao, "Sentence group research", (**1984**), Zhejiang Education Press.
[15] J. Cohen, "A coefficient of agreement for nominal scales", Educational and Psychological Measurement, vol. 20, no. 1, (**1960**), pp. 37-46.
[16] A. K. McCallum, "Mallet: a machine learning for language toolkit", (**2002**) http://mallet.cs. umass.edu.

Authors

Yancui Li, She received the Master degree in computer science and technology from Soochow University, China, in 2008. Now she is a Ph.D. candidate of Soochow University. She works as lecturer in Henan Institute of Science and Technology since 2008. His main research interests include natural language processing and data mining.

Hongyu Feng, she received the Master degree in computer science and technology from South West Jiaotong University, China in 2006. She now works in Henan institute of Science and Technology. The author's major field of study is Intelligent computing, natural language processing and computer application.

Wenhe Feng, He received the Ph.D. degree in linguistics from Wuhan University, China in 2011. Currently, he is a postdoctoral fellow at School of Computer, Wuhan University, China. His research interests include natural language processing and Chinese linguistics.