Maximum Entropy for Chinese Comma Classification with Rich Linguistic Features

Xiaojuan Li
School of Mathematics and Computer Science, Guizhou Normal University
596025763@qq.com

Hua Yang*
School of Mathematics and Computer Science, Guizhou Normal University
College of Chinese Language and Literature, Wuhan University
yanghuastory@foxmail.com

JiangPing Huang
School of Computer, Wuhan University
hjp@whu.edu.cn

Abstract

Discourse relation is an important content of discourse semantic analysis, and the study of punctuation is of importance for discourse relation. In this paper, we propose a method of Chinese comma classification based on maximum entropy (ME). This method classifies the sentence relation based on comma with ME by extracting rich linguistic features before and after the commas in sentences. Experimental results show that this method of sentence relation based on comma is feasible.

1 Introduction

Discourse consists of word, phrase, sentence and sentence group, also known as text or utterance. Discourse relation studies the intrinsic structure of natural language text and understands the semantic relation between the text units, which plays a vital role in language understanding and natural language generation, is a challenge and difficult research hotspot in recent years((Li Yan-cui et al.,2013). Discourse relation is a fundamental work in the research of discourse analysis. Discourse relation means the logical semantic relation, between two text unit (section, clause, sentence, sentence group, paragraphs, etc.) in one discourse, such as coordinative relation, progressive relation, adversative relation (Sun Jing et al., 2014), etc. Defining a hierarchical semantic relationship type system to extend sentence semantic analysis results in that discourse level of semantic information become one of the important ways to solve the discourse semantic analysis, which is benefit to many NLP tasks such as automatic summarization, automatic question answering and machine translation (Zhang Mu-yu et al., 2013).

The commas separates a sentence into two parts, each part is called an argument of the sentence. Discourse relation can be generally classified into explicit relation and implicit relation. Explicit relation recognition is to identify the logical relationship between two arguments in the presence of conjunctions (Sun Jing et al., 2014) while implicit relation recognition is to identify the logical relationship without the presence of conjunctions. Example 1 exemplifies the explicit relation of coordination with the conjunction word "并(and)", and example 2 exemplifies the implicit relation of coordination in the absence of "并(and)", in which conjunction does not appear. For the implicit relation recognition, the absence of conjunction entails methods that can deduce the semantic type from other features in the context before and/or after commas. In previous researches, explicit relation recognition often has a higher precision only based on conjunction, while implicit relation recognition is much more difficult than explicit relation recognition. Some additional information is gradually introduced in addition to lexical features (Zhang Mu-yu et al., 2013).

eg. 1：跳水选手已全部抵达罗马，**并**开始赛前训练。

"All divers have arrived in Rome, and start training before the game."

eg. 2：中国的稳定和发展有利于世界的和平与发展，中国的繁荣与稳定是澳门繁荣与稳定的根本保证。

"China's stability and development are conducive to world's peace and development, China's prosperity and stability are the fundamental guarantee of Macro's prosperity and stability."

Most researches about discourse relation recognition are mainly for English. Although there are some Chinese-oriented research (Jin Mei-xun et al., 2004; Xu Sheng-qin and Li Pei-feng, 2013; Yang ya-qin and Xue Nianwen, 2012), they are mainly concentrated on the analysis and corpus annotation, rarely involving discourse relation recognition; and existing research mostly directly used the English discourse relation system, ignoring the linguistic characteristics of Chinese language itself.

According to the classification of compound sentence theories (Xing Fu-yi, 2001; Lv Shuxiang and

Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 11–17,
Wuhan, China, 20-21 October 2014

Zhu De-xi, 1952; Shao Jing-min, 2007), in this paper, we propose 9 categories of Chinese comma classification for sentence relation, including Coordination(并列), Interpretation(阐释), Location(地点), Progressiveness(递进), Reliance(凭借), Subsequence(顺承), Time(时间), Purpose(目的), Cause and Effect(因果), and classify Chinese comma into these 9 classes with maximum entropy method (ME), the corpus we used is annotated with a well-established representation scheme for Chinese comma, and the features we used are extracted from the corpus that is based on the sentences' words information on both sides of the comma. We carried out the classification experiment on both the explicit relation recognition and the implicit relation recognition respectively consisted of the 9 categories mentioned above.

The rest of the paper is organized as follows. In section 2, we describe the related work about comma classification research. Section 3 introduces the features we used and other features selecting method used in related work. Section 4 reviews ME method and describe the comma classification method based on ME model. In section 5, we present the process of our experiment and evaluate the experimental result. In section 6, we analyze the causes that lead to the main classification error in different aspects. Finally, a conclusion and future work are put forward.

2 Related Work

As elemental segmentation units of discourse, punctuations provide a new clue for discourse analysis. Many researches about punctuation are closely related with many natural language processing tasks, such as long sentences segmentation, elementary discourse unit recognition, the classification of the relationship between sentences, semantic disambiguation, etc. 16 kinds of punctuations are widely used in Chinese, such as comma, period, question mark, etc. With more than 20 different usages, comma is one of the most common punctuations. Chinese comma can be used to separate coordinate composition or coordinate clause of the sentence, or to separate the words, phrases, clauses which indicate time, place, purpose, condition, or to express a pause between the clauses separated by conjunction (Gu Jing-jing and Zhou Guo-dong, 2014), etc. In recent years, with the progress of the research about punctuation, the study of comma classification gradually caught attention.

Jin and Li (Jin Mei-xun et al., 2004) viewed comma as an important role in long Chinese sentence segmentation, they proposed a method for classifying commas in Chinese sentences by their context, then segmented a long sentence according to the classification results. Element discourse unit (EDU) recognition is a fundamental task of discourse analysis and Chinese punctuation is viewed as a elementary delimiter. Xu Sheng-qin and Li Pei-feng (Xu Sheng-qin and Li Pei-feng, 2013) considered Chinese comma to be the boundary of the discourse units and anchor discourse relations between units separated by comma. They classified comma's role into seven major types and implemented automatic disambiguation of the Chinese comma type. Xue and Yang (Xue Nian-wen and Yang Ya-qin, 2011) held that the central problem of Chinese sentence segmentation was comma disambiguation, and in some context it identifies the boundary of a sentence just as a period, a question mark, or an exclamation mark does. Yang and Xue (Yang ya-qin and Xue Nian-wen, 2012) further pointed out that the Chinese comma signifies the boundary of discourse units and also anchors discourse relations between adjacent text spans, and they proposed a discourse structure-oriented classification of the comma that can be automatically extracted from the Chinese Treebank based on syntactic patterns, and use this method to disambiguate the Chinese comma.

In this paper, we propose a method of sentence relation classification based on rich linguistic features around Chinese comma in sentences. We try to find out the difference among sentence relation types by rich linguistic features, which is found by potential semantic rules derived by statistical method, which is of significance especially for the implicit relation recognition.

3 Features Selection

Currently, few research about sentence relation is based on comma. Sun jing (Sun Jing et al., 2014) classified the discourse relation into four categories: cause and effect(因果), coordination(并列), transition(转折), explanation(解说) with maximum entropy, on the basis of utilizing a set of context features, lexical features and dependency tree features extracted from the corpus of Chinese discourse built by themselves. Lin (Lin Zi-heng et al., 2009) implemented an implicit discourse relation classifier and showed initial results based on the recently released Penn Discourse Treebank. The features they used include the modeling of the context of relations, features extracted from constituent parse trees and dependency parse trees, and word pair features. Zheng (Zheng Lue-xing et al., 2013) presented an approach of Chinese coordination relations recognition based on CRFs. They extracted role information according to their functions in the generation of Chinese coordination relations.

We analyze the feature of different types of sentences, refer to the features proposed in the paper of Li Yancui (Li Yan-cui et al., 2013) and Xue (Xue Nian-wen and Yang Ya-qin, 2011), and propose to learn discourse relation rules through linguistic features of the sentences. This method extract linguistic features from both sides of comma in the sentence. Before extracting the features, the following pre-processing is adopted: 1) segment the sentences into words by using the Chinese lexical analysis

system (ICTCLAS) designed by institute of computing technology, Chinese academy of sciences; 2) eliminate the extremely precise POS type for the words, which belongs to the same POS on more general level. For example, "nr" expresses name, "ns" expresses place name), and we use "n" to express the noun, "v" to express verb uniformly, etc.

We call the sentence on the left of the comma as argument 1, denoted as "*l*", and call the sentence after the comma as argument 2 and express it with "*r*". Features we selected and their descriptions are shown in table 1.

Table 1 the selected features and their description

	feature	description
1	f1,f1_p	The first word of argument 1 and its part of speech(POS)
2	f2	Conjunction that connects the clauses on both sides of the comma, if no conjunction appear, f2 =null
3	f3	Difference of clause lengths between argument 1 and argument 2, if the length of argument 1 is greater than the argument 2, f3=1, otherwise f3=0
4	f4,f4_p	The first word of argument 2 and its POS
5	f5_l,f5_r	Whether the *l* and *r* contain a conjunction
6	f6,f6_p	The last word of argument 1 and its POS
7	f7	The POS of the first word combination of argument 1 and argument 2(f1_p+f4_p)
8	f8	Combination of the POS of the first word and last word in argument 1(f1_p+f6_p)
9	f9	Let *x* denote whether the first word of *l* is a conjunction, *x=1* if the first word of *l* is a conjunction, else *x =0*. f9 is the combination of *x* and *POS* of the first word of *l*
10	f10	Feature 10 is analogous to f9, while x denotes whether the last word of *l* is a conjunction.
11	f11	Feature 11 is analogous to f9, while x denotes whether the first word of *r* is a conjunction.
12	f12	f12=1 if the first word and the last word of argument 1 constitute a conjunction, else f12=0

Features of case 1 and case 2 mentioned above are as follows.

1：f1=跳水选手, f1p=n, f2=并, f3=1, f4=并, f4p=c, f5l=0, f5r=1, f6=罗马, f6p=n, f7=n+c, f8=n+n, f9=0+n, f10=0+n, f11=1+c, f12=0

2：f1=跳水选手, f1p=n, f2=null, f3=1, f4=开始, f4p=ad, f5l=0, f5r=0, f6=罗马, f6p=n, f7=n+v, f8=n+n, f9=0+n, f10=0+n, f11=1+v, f12=0

4 Maximum Entropy for Comma Classification

Maximum entropy model (ME) method is to select the model with the maximum entropy that meets some constraint conditions. Maximum entropy model can be applied to classification(Li Hang, 2012, Sang Haiyan et al., 2013).

In our implementation, ME model uses the features listed in table1.

Let C be the set of types of the 9 sentence relation classes we have defined, and *S be the* sentence set, we can calculate $p(c_j \mid s_i)$ through maximum entropy model, which means the probability s_i belongs to c_j, where $s_i \in$S and $c_j \in$C. For comma classification problem, c_j with $\arg\max p(c_j \mid s_i)$ will be the class that the sentence s_i belongs to.

The comma classification method is similar to text classification method, their basic idea is to use learning set composed of training samples to train a classifier, to test the performance of the classifier with testing samples in testing set, and use the trained classifier to classify new sentences.

5 Experiments and Evaluation

Corpus used in our experiment is rebuilt from part of CTB 5.0. We annotated it with the information of class. The corpus is divided into explicit relation and implicit relation according to whether the sentences contain conjunction. The distribution of the sample set for each class is shown in table 2 .

The eigenvector expressed with features in Table 1 for each sentence in Table 2 is obtained. All the eigenvectors obtained constitute our data set. The data set is divided into training data set and testing data set with the proportion of 80% : 20%, 10-times 10-fold cross-validation policy is employed. All of above prepared, one of the mallet toolkit classifier--maximum entropy (MaxEnt) classifier is adopted to train and test the final model. The experimental results, i.e., classification precisions for all sentence relation class, are shown in table 3.

Table 2 distribution of sentence relationship

number	categories	data	
		explicit	implicit
1	Coordination(BL)	25	24
2	Interpretation(CS)	25	25
3	Location(DD)	25	6
4	Progressiveness(DJ)	25	11
5	Reliance(PJ)	25	10
6	Subsequence(SC)	12	25
7	Time(SJ)	25	24
8	Purpose(MD)	25	6
9	Cause and Effect(YG)	25	25

We conducted several experiments on different training set size and testing set size. Results show that the unbalance of training set size has a significant effect on the experimental results. So we use the same training set size avoid this instability. As can be seen in table 3, results for four relations (Location, Progressiveness, Reliance and Purpose) are absent. The reason for the absence is that the corresponding precision is unreliable due to the sparseness of related samples in training data showed in Table 2. In addition, the precision for implicit relations is significantly lower than that for the explicit relations.

Table 3 experimental results

category of relationship	explicit precision	implicit precision
Coordination(BL)	56.5%	49.7%
Interpretation(CS)	62.4%	47.3%
Location(DD)	84.9%	--
Progressiveness(DJ)	63.2%	--
Reliance(PJ)	71.2%	--
Subsequence(SC)	--	38.9%
Time(SJ)	43.1%	54.2%
Purpose(MD)	55.5%	--
Cause and Effect(YG)	72%	74.1%
ALL	65.2%	50.6%

6 Analysis

Table 4 shows the details of explicit relation classification, which includes the percentage of the samples that are correctly classified and falsely classified into other classes. Each item in Table 4 is the average calculated from 10 times repeated experiment. Table 5 is corresponding result for implicit relation classification.

As can be seen in Table 4 and Table 5, main errors mainly occur as follows:

（1）For explicit relation, Many Location relation and Time relation are falsely classified into each other; Time relation is cline to be classified into Reliance; Purpose relation is classified into Reliance. The reasons for falsely classification for Location and Time is: the first word in argument 1 is preposition in most cases, and the last word in argument 1 means a location expressed as “f” in some cases, as shown in example 3 and example 4; for the relation of Purpose and Reliance, the reason for falsely classification is that the first word in argument 1 is preposition in most cases, as shown in example 5 and example 6 ; for the relation of Time and Reliance, the reason for falsely classification is that the first word in argument 1 is preposition in most cases and their conjunction is composed of the first word and the last word of argument 1, as shown in example 4 and example 6.

Table 4 details for explicit relation classification

	Interpre-tation	Location	Progres-siveness	Reli-ance	Time	Coordina-tion	Purpos e	Cause and Effect
Interpre-tation	64%	4%	11%	0	0	14%	4%	4%
Location	0	73%	5%	0	14%	0	5%	5%
Progres-siveness	15%	7%	56%	0	0	22%	0	0
Reliance	4%	0	0	83%	4%	0	8%	0
Time	0	28%	8%	12%	44%	4%	4%	0
Coordi-nation	12%	0	16%	0	0	64%	8%	0
Purpose	5%	0	0	20%	5%	5%	65%	0
Cause and Ef-fect	0	0	7%	0	0	0	0	93%

Table5 details for implicit relation classification

	Coordination	Interpretation	Subsequence	Time	Cause and Effect
Coordination	12%	23%	35%	31%	0
Interpretation	37%	56%	7%	0	0
Subsequence	0	6%	71%	24%	0
Time	21%	8%	33%	33%	4%
Cause and Effect	0	4%	0	0	96%

eg. 3：**在**今天的比赛**中**，中国国际大师徐俊迎战队友、国际特级大师叶荣光。（地点）

"In today's competition, the Chinese international master Jun xu will meet his teammate who is an international grandmaster Rongguang Ye."

eg. 4：**在**这一巨大的变革**中**，德国成为最大的得益者。（时间）

"In this huge change, Germany is the biggest beneficiary."

eg. 5：**为**解决庞大资金需求，公司正争取发行股票和尝试更多的融资渠道。（目的）

"To solve the large capital demand, the company is seeking to issue shares and try more financing channels."

eg. 6：**据预测**，今年全球经济增长幅度可达到百分之四点一。(凭借)

"It is predicted that the global economic growth can reach 4.1% this year."

Example 3, 4, 5, 6 represents the Location, Time, Purpose and Reliance respectively. In example 3, the conjunction is the combination of “在” and “中”, and the pos-of-part of “在” is preposition, the “中” means location. In example 4 sentence, the conjunction is the combination of “在” and “中”, the pos-of-part of this conjunction is same as example 3. In example 5 sentence, the conjunction is “为”, and its pos-of-part is preposition. In example 6 sentence, the conjunction is the combination of “据” and “预测”, the pos-of-part of “据” is preposition.

（2）Subsequence and other relations class in implicit relations

Implicit relation has no obvious semantic type sign (conjunction) so that it is difficult to determine the existence of relation and the relation type without human's judgment. Subsequence relation is very special that can not be easily differentiated from other relation types even by human, which often result in controversy among annotator, and reduce precision of the implicit relation recognition. For example, the subsequence relation expresses the sentence relation of time, space or logical sequence, etc. However, most other relations involve certain subsequence relation to some degree, resulting in that other relation is easily classified as subsequence in the implicit relation recognition. Example 7 represents the coordination, and example 8 represents the subsequence as shown below.

eg. 7：拉美是一个充满希望的大陆，具有巨大的发展潜力。（并列）

"Latin America is a continent of hope, possessing huge development potential."

eg. 8：《新中东》一书原为英文版，去年秋冬之交出版。（顺承）

"“*The new Middle East*” *was* English version, and published since the turn of the last autumn and winter."

（3）Coordination, Progressiveness and Interpretation, Coordination and Time

In Chinese, Coordination relation describes the parallelism between clauses or words, which can be split into two independent arguments by the comma. Progressiveness relation always implies that the second argument contains more information. However, in many cases, the conjunction “并” (expressing parallelism in most cases) can also express progressive relationship. No matter in the explicit or implicit relationship recognition, Progressive and Coordinate are easy to be confused with each other because they have similar structure and POS information.

The examples below are two sentences extracted from the corpus, example 9 represents the coordination, and example 10 represents the progressiveness.

eg. 9：两年多来两国经贸合作已顺利起步，并取得可观的进展。（并列）

"For more than two years the bilateral economic and trade cooperation has started smoothly, and achieved considerable progress. "

eg. 10：中国已确定了未来五年高技术研究重点，并着手制订下世纪的高科技研究计划。（递进）

"China has determined the high-tech research focal point of the next five years, and has began to make plan of high-tech research for next century."

It is difficult to analyze the difference between coordination and progressiveness from above examples, which is one of the causes in classification errors.

7 Conclusions and future work

We proposed the Chinese comma classification based on Chinese discourse relationship corpus. Rich linguistic features have been selected in the classification and sentence relations are classified into 9 categories with maximum entropy method. The experimental results show that the method based on linguistic features for classification of comma is feasible. However, from the result we can see that the overall classification precision still needs to be improved, especially for the implicit relation. In future work, we will further study how to extract more effective features, try to attach great importance to the role of conjunctions, which is vital to distinguish the explicit relation between sentences, and combine these features with the structure of the sentences to improve classification accuracy. In addition, we also need to solve the problem of the small scale of sample set and data sparsity.

Acknowledgement

This paper is supported by Natural Science Foundation Project (61070243, 6133012), Major Project of Invitation for Bid of National Social Science Foundation (11&ZD189), Guizhou High-level Talent Research Project (TZJF-2010-048), Guizhou Normal University PhD Start-up Research Project (11904-05032110011), and Governor Special Fund Grant of Guizhou Province for Prominent Science and Technology Talents (identification serial number "黔省专合字(2012)155号"), China Postdoctoral Science Foundation(2013M531730).

Reference

Li Yan-cui, Feng Wen-he, Zhou Guo-dong. 2013. *Research of Chinese clause identificiton based on comma*. Journal of Beijing University (Natural Science Edition), 2013(01): 7-14.

Sun Jing, Li Yan-cui, Zhou Guo-dong. 2014. *Research of Chinese implicit discourse relation recognition.* Journal of Beijing University (Natural Science Edition), 2014(01): 111-117.

Gu Jing-jing, Zhou Guo-dong. *Chinese comma classification based on segmentation and part of speech-tagging*. 2014. Computer Engineering and Applications.

Jin Mei-xun, Mi-Young Kim, Donggil Kim, Jong-Hyeok Lee. 2004. *Segmentation of Chinese long-sentences using commas*. SIGHAN2004

Xu Sheng-qin, Li Pei-feng. 2013. *Recognizing Chinese elementary discourse unit on comma*. Asian Language Processing (IALP),2013 Internationa. IEEE, 2013: 3-6.

Yang Ya-qin , Xue Nian-wen . 2012. *Chinese comma disambiguation for discourse analysis*. Proceedings of the 50th Annual Meeting of the Asso. Association for Computational Li, 2012: 786-794.

Xing Fu-yi. 2001. *The study of Chinese complex sentence*. Beijing:Commercial Press,2001

Lv Shu-xiang, Dexi Zhu. 1952. *Grammatical rhetoric speech*. Liaoning:Liaoning Education Press,1952

Shao Jing-min. 2007. *The general theory of modern Chinese*. Liaoning:Shanghai Education Press,2007

Lin Zi-heng, Kan Min-yen, Hwee Tou Ng . 2009. *Recognizing implicit discourse relations in the Penn Discourse Treebank*. Proceedings of the 2009 Conference on Empirical Me Association for Computational Li, 2009: 343-351.

Li Hang. 2012. *Statistical learning method*. Beijing: Tsinghua University Press, 2012: 80-87

Sang Hai-yan, Gu Lia-Altenbek, Niu Ning-ning. 2013. *Kazakh part-of-speech tagging method based on maximum entropy.* Computer Engineering and Applications, 2013, 49(11): 126-129,16.

Zhang Mu-yu,Song Yuan,Qin Bing,Liu Ting 2013. *Chinese Discourse Relation Recognition.* Journal of Chinese information, 2013,27(6): 51-57.

Xue Nian-wen, Yang Ya-qin . 2011. *Chinese sentence segmentation as comma classification*. Proceedings

of the 49th Annual Meeting of the Asso. Association for Computational Li, 2011: 631-635.

Zheng Lue-xing, Lv Xue-qiang, Liu Kun, Lin Jin. 2013. *Automatic Identification of Chinese Coordination Relations*. Journal of Beijing University (Natural Science Edition),2013,49(1):20-24.