Enhancement of a Chinese Discourse Marker Tagger with C4.5

Benjamin K. T'sou[1], Tom B. Y. Lai[2], Samuel W. K. Chan[3], Weijun Gao[4], Xuegang Zhan[5]

[1 2 3]Language Information Sciences Research Centre
City University of Hong Kong
Tat Chee Avenue, Kowloon
Hong Kong SAR, China

Northeastern University, China

{[1]rlbtsou, [2]cttomlai}@uxmail.cityu.edu.hk, [3]swkchan@cs.cityu.edu.hk,
[4]wjgao@mail.neu.edu.cn, [5]zxg@ics.cs.neu.edu.cn

Abstract

Discourse markers are complex discontinuous linguistic expressions which are used to explicitly signal the discourse structure of a text. This paper describes efforts to improve an automatic tagging system which identifies and classifies discourse markers in Chinese texts by applying machine learning (ML) to the disambiguation of discourse markers, as an integral part of automatic text summarization via rhetorical structure. Encouraging results are reported.

Keywords: discourse marker, Chinese corpus, rhetorical relation, automatic tagging, machine learning

1 Introduction

Discourse refers to any form of language-based communication involving multiple sentences or utterances. The most important forms of discourse of interest to Natural Language Processing (NLP) are text and dialogue. The function of discourse analysis is to divide a text into discourse segments, and to recognize and re-construct the discourse structure of the text as intended by its author.

Automatic text abstraction has received considerable attention (Paice 1990). Various systems have been developed (Chan et al. 2000). Ono et al. (1994), T'sou et al. (1992) and Marcu (1997) focus on discourse structure in summarization using the Rhetorical Structure Theory (RST, Mann and Thompson 1986). The theory has been exploited in a number of computational systems (e.g. Hovy 1993). The main idea is to build a discourse tree where each node of the tree represents an RST relation. Summarization is achieved by trimming unimportant sentences on the basis of the relative saliency or rhetorical relations.

The SIFAS (Syntactic Marker based Full-Text Abstraction System) system has been implemented to use discourse markers in the automatic summarization of Chinese (T'sou et al. 1999). In this paper, we report our efforts to improve the SIFAS tagging system by applying machine learning techniques to disambiguation of discourse markers. C4.5 (Quinlan, 1993) is used in our system.

2 Manual Tagging Process

To tag the discourse markers, the following coding scheme is designed to encode *Real* Discourse Markers (RDM) appearing in the SIFAS corpus (T'sou et al. 1998). We describe the i^{th} discourse marker with a 7-tuple RDM_i:

RDM_i=< DM_i , RR_i , RP_i , CT_i , MN_i , RN_i , OT_i >, where

DM_i : the lexical item of the *D*iscourse *M*arker, or the value '*NULL*'.
RR_i : the *R*hetorical *R*elation in which DM_i is a constituent marker.
RP_i : the *R*elative *P*osition of DM_i.
CT_i : the *C*onnection *T*ype of RR_i.
MN_i : the Discourse *M*arker Sequence *N*umber.
RN_i : the Rhetorical *R*elation Sequence *N*umber.
OT_i : the *O*rder *T*ype of RR_i. The value of OT_i can be 1, -1 or 0, denoting respectively the normal order, reverse order or irrelevance of the premise-consequence ordering of RR_i.

For *apparent* discourse markers that do not function as a real discourse marker in a text, a different coding scheme is used to encode them. We describe the i^{th} apparent discourse marker using a 3-Tuple ADM_i:

ADM_i=< LI_i, *, SN_i >, where

LI_i : the *L*exical *I*tem of the apparent discourse marker.
SN_i : the *S*equence *N*umber of the apparent discourse marker.

In Chinese, discourse markers can be either words or phrases. To tag the SIFAS corpus, all discourse markers are organized into a discourse marker pair-rhetorical relation correspondence table. Part of the table is shown Table 1.

To construct an automatic tagging system, let us first examine the sequential steps in the tagging process of a human tagger.

S1. Written Chinese consists of running texts without word delimiters; the first step is is to segment the text into Chinese word sequences.
S2. On the basis of a discourse marker list, we identify those words in the text which appear on the list as Candidate Discourse Markers (CDMs).
S3. To winnow Real Discourse Markers (RDMs) and Apparent Discourse Markers (ADMs) from the CDMs, and encode the ADMs with a 3-tuple.
S4. To encode the RDM with a 7-tuple according to a Discourse Marker Pair-Rhetorical Relation correspondence table.

Relation	Front	Back	Connection Type	Order Type
Adversativity		可是	Inter	1
Adversativity		可是	Intra	1
Causality	因为	所以	Intra	1
Causality	之所以		Intra	-1
.				
.				

Table 1 Discourse Marker Pair-Rhetorical Relation Table

3 Automatic Tagging Process

The identification of candidate discourse markers is based on a discourse marker list, which now contains 306 discourse markers plus a NULL marker. The markers are extracted from newspaper editorials of Hong Kong, Mainland China, Taiwan and Singapore. These markers constitute 480 distinct discontinuous pairs that correspond to 25 rhetorical relations. In actual usage, some discourse marker pairs designate multiple rhetorical relations according to context. Some pairs can represent both INTER-sentence and INTRA-sentence relations. Thus the correspondence between the discourse marker pairs and the rhetorical relations is not single-valued. Some discourse marker pairs correspond to more than one rhetorical relation or connection type. We have 504 correspondences between the discourse marker pairs and the rhetorical relations.

In practice, one discontinuous constituent member of a marker pair is often omitted. We use the NULL marker to indicate the omission. In the 504 correspondences, 244 of them are double constituent marker pairs, 260 are single constituent markers (i.e. One of the markers is NULL). And in the 244 double constituent markers, only 3 are not single-valued correspondences (one of which is an INTER/INTRA relation, and can easily be distinguished.). Thus the tagging of the 244 double constituent markers is basically a table searching process. But for the 260 single constituent markers, the identity of the NULL marker is often difficult to determine.

The SIFAS tagging system works in two modes: automatic and interactive (semi-automatic). The automatic tagging procedure is as follows:

1. Data preparation: Input data files are modified according to the required format.
2. Word segmentation: Because there are no delimiters between Chinese words in a text, words have to be extracted through a segmentation process.
3. CDM identification
4. Full-Marker RDM recognition
5. ADM identification (first pass, deterministic)
6. CDM feature extraction
7. ADM identification (2nd pass, via ML)
8. Tagging NULL-marker CDM pairs (via ML)
9. ADM and RDM sequencing, proof-reading, training data generation, and statistics

The following principles are adopted by the tagging algorithm to resolve ambiguity in the process of matching discontinuous discourse markers:

1. ***the principle of greediness***: When matching a pair of discourse markers for a rhetorical relation, priority is given to the first matched relation from the left.
2. ***the principle of locality***: When matching a pair of discourse markers for a rhetorical relation, priority is given to the relation where the distance between its constituent markers is shortest.
3. ***the principle of explicitness***: When matching a pair of discourse markers for a rhetorical relation, priority is given to the relation where both markers are explicitly presented.
4. ***the principle of superiority***: When matching a pair of discourse markers for a rhetorical relation, priority is given to the inter-sentence relation whose back discourse marker matched with the first word of a sentence.
5. ***the principle of Back-marker preference***: This is applicable only to rhetorical relations where either the front or the back marker is absent, or to a NULL marker. In such cases, priority is given to the relation with the back marker present.

Steps 1 to 6 and the five principles underlie the original naïve tagger of the SIFAS system (T'sou et al. 1998), which also contains the system framework.

4 Improvement

4.1 Problems

Many Chinese discourse markers have both discourse senses and alternate sentential senses in different context. For a human tagger, steps S3 and S4 in section 2 are not difficult because he/she can identify an ADM/RDM based on his/her text comprehension. However, for an automatic process, it is quite difficult to distinguish an ADM from an RDM if no syntactic/semantic information is available.

Another problem is the location of NULL-Marker described above. Our earlier statistics showed some characteristics in the distance measured by punctuation marks. Statistics from 80 tagged editorials show that most of the relations are INTRA-Sentence relations (about 93%), about 70% of the INTRA RDM pairs have NULL markers. Most of these RDM pairs are separated by ONE comma (62%). These statistics show the importance of the problems of positioning the NULL markers.

The naïve tagger partially solved the CDM discrimination and NULL marker location problems. Our experiment shows that about 45% of the ADMs can be correctly identified, and about 60% of the NULL markers can be correctly located one comma/period away from the current RDM. This leaves much room for improvement.

One solution is to add a few rules according to previous statistics. The original naïve tagger did not assume any knowledge of the statistics and behavioral patterns of discourse markers. From the error analysis, we extracted some additional rules to guide the classification and matching of the discourse markers. For example, one of the rules we extracted is:

> "A matching pair must be separated by at least two words or by punctuation marks". Using this rule, the following full marker matching error is avoided.

<抓紧><实施><退><耕><*还*,conjunction,Front,Intra,5,5,1><林><*还*, conjunction, Back, Intra, 6,5,1><草>、<治理><水土流失><等><生态><建设>，<是,*,7><西部><地区><今后><<需要><着力><完成><的><重大><任务>。

Another solution is to use syntactic/semantic information through machine learning.

4.2 C4.5

Most empirical learning systems are given a set of pre-classified cases, each described by a vector of attribute values, and construct from them a mapping from attribute values to classes. C4.5 is one such system that learns decision-tree classifiers. It uses a divide-and-conquer approach to growing decision trees. The current version of C4.5 is C5.0 for Unix and See5 for Windows.

Let attributes be denoted $\boldsymbol{A=\{a_1, a_2, ..., a_m\}}$, cases be denoted $\boldsymbol{D=\{d_1, d_2, ..., d_n\}}$, and classes be denoted $\boldsymbol{C=\{c_1, c_2, ..., c_k\}}$. For a set of cases $\boldsymbol{D}$, a test $\boldsymbol{T_i}$ is a split of $\boldsymbol{D}$ based on attribute $\boldsymbol{a_i}$. It splits $\boldsymbol{D}$ into mutually exclusive subsets $\boldsymbol{D_1, D_2, ..., D_p}$. These subsets of cases are single-class collections of cases.

If a test $\boldsymbol{T}$ is chosen, the decision tree for $\boldsymbol{D}$ consists of a node identifying the test $\boldsymbol{T}$, and one branch for each possible subset $\boldsymbol{D_i}$. For each subset $\boldsymbol{D_i}$, a new test is then chosen for further split. If $\boldsymbol{D_i}$ satisfies a stopping criterion, the tree for $\boldsymbol{D_i}$ is a leaf associated with the most frequent class in $\boldsymbol{D_i}$. One reason for stopping is that cases in $\boldsymbol{D_i}$ belong to one class.

C4.5 uses ***arg max(gain(D,T))*** or ***arg max(gain ratio(D,T))*** to choose tests for split:

$$Info(D) = -\sum_{i=1}^{k} p(c_i, D) * \log 2(p(c_i, D))$$

$$Split(D,T) = -\sum_{i=1}^{p} \frac{|D_i|}{|D|} * \log 2(\frac{|D_i|}{|D|})$$

$$Gain(D,T) = Info(D) - \sum_{i=1}^{p} \frac{|D_i|}{|D|} * Info(D_i)$$

$$Gain\,ratio(D,T) = gain(D,T) / Split(D,T)$$

where, $\boldsymbol{p(c_i,D)}$ denotes the proportion of cases in $\boldsymbol{D}$ that belong to the ith class.

4.3 Application of C4.5

Since using semantic information requires a comprehensive thesaurus, which is unavailable at present, we only use syntactic information through machine learning.

The attributes used in the original SIFAS system include the candidate discourse marker itself, two words immediately to the left of the CDM, and two words immediately to the right of the CDM. The attribute names are F2, F1, CDM, B1, B2, respectively (T'sou et al, 1999). SIFAS only uses the Part Of Speech attribute of the neighboring words. This reflects to some degree the syntactic characteristics of the CDM.

To reflect the distance characteristics, we add two other attributes: the number of discourse delimiters (commas, semicolons for INTRA-sentence relation, periods and exclamation marks for INTER-sentence relation) before and after the current CDM, denoted Fcom and Bcom, respectively. For the location of the NULL marker, we still add an actual number of delimiters Acom.

The order of these attributes is: CDM, F1, F2, B1, B2, Fcom, Bcom Acom for Null marker location, and CDM, F1, F2, B1, B2, Fcom, Bcom, IsRDM for CDM classification, where IsRDM is a Boolean value.

The following are two examples of cases:

并且,?,q,a,a,7,1,1 for NULL marker location

包括,d,?,u,?,1,0,F for CDM classificati on

where "?" denotes that no corresponding word is at the position (beginning or end of sentence); a, d, q, and u are part-of-speech symbols in our segmentation dictionary, representing adjective, adverb, classifier, and auxiliary, respectively.

The following are two examples of the rules generated by the C4.5. The first is a CDM classification rule, and the other is a NULL marker location rule.

Rule 5: (11/1, lift 2.2)
CDM = 并
B1 = v
Fcom > 0
➔ class T [0.846]

which can be explained as: if the word after the CDM "并" is a verb, and there is one comma in the sentence, before "并", then "并" is an RDM.

Rule 22: (1, lift 3.4)
B2 = p
Fcom > 1
➔ class 2 [0.667]

which can be explained as: if the second word after the RDM is a preposition, and there is more then one commas before the current RDM, then the location of the NULL marker is two commas away from the RDM.

4.4 Objects in the SIFAS system

The objects in the new SIFAS tagging system are listed below.

1. Dictionary Editor: for the update of word segmentation dictionary and the rhetorical relation table.
2. Data Manager: for the modification of the input data (editorial texts) to conform with the required format.
3. Word Segmenter: for the segmentation of the original texts, and the recognition of CDMs.
4. RDM Tagger: The initial identification of RDMs is a table searching process. All those full-marker pairs are identified as rhetorical relations according to the principles described above. For those Null-marker pairs, the location of the Null maker is left to the rule interpreter.
5. ADM Tagger: The identification of ADMs is also a table searching process, because, without other syntactic/semantic information, the only way to identify ADMs from the CDMs is to find out that the CDM cannot form a valid pair with any other CDMs (including the NULL marker) to correspond to a rhetorical relation.
6. CDM Feature Extractor: For those untagged CDMs, the classification is carried out through C4.5. The Feature Extractor extracts syntactic information about the current CDM and send it to the Rule Interpreter (see below).
7. Rule Interpreter: C4.5 takes feature data file as the input to construct a classifier, and the rules formed are stored in an output file. The rule interpreter reads this output file and applies the rules to classify the CDMs. In our system, The Rule Interpreter functions as a NULL Marker Locator and a CDM classifier.
8. Sequencer: for the rearrangement of RDM and ADM order number. In the rearranging process, the Sequencer also extracts statistical information for analysis.
9. Interaction Recorder: for the recording of user interaction information for

statistics use.

10. Data Retriever: for data retrieval and browsing.

5 Evaluation

In order to evaluate the effectiveness of the tagging system in terms of the percentage of discourse markers that can be tagged correctly, we have chosen 80 tagged editorials from Ming Pao, a Chinese newspaper of Hong Kong, in the duration from December 1995 to January 1996 to form a training data set. Then we randomly selected 20 editorials from Mainland China and Hong Kong newspapers for the system to tag automatically, and then manually checked the results.

The total CDMs in the training data set is 4764, in which 2116 are RDMs and 2648 are ADMs. The distribution of INTER-sentence relations, INTRA-sentence relations, and NULL marker pairs is shown below.

Total Relations	Inter-Sentence Relations	Intra-Sentence Relations	Relations with NULL marker pair
1589	98	1491	1062
100%	6.17%	93.83%	66.83%

Table 2 Distribution of INTER-/INTRA-sentence relations, and NULL marker pairs

Our evaluation is based on counting the number of discourse markers that are correctly and incorrectly tagged.

The total CDMs in the test data set is 1134, in which 563 are RDMs and 571 are ADMs. The distribution of INTER-sentence relations, INTRA-sentence relations, and NULL marker pairs in the test data set is shown in Table 3.

Total Relations	Inter-Sentence Relations	Intra-Sentence Relations	Relations with NULL marker pair
424	23	401	285
100%	5.42%	94.58%	67.22%

Table 3 Distribution of INTER-/INTRA-sentence relations, and NULL marker pairs in testing data set

Tagged Relations	Correctly tagged	NULL marker	Full marker	ADM/R DM	Other errors
451	399	11	1	65	3

Table 4 Test Results

From the test results shown in Table 4, we can see that most of the errors are caused by the misclassification of the CDMs. An example of *Other errors* is shown below. The following sentence is from an editorial of People's Daily.

<我们><再次><重申>，<*NULL*,sufficiency,Front,Intra,0,81,1><一个中国><的><原则><不容><回避>，<不><承认><一个中国><的><原则>、<不><承认><台湾><是,*,80><中国><的><一部分>，<台湾海峡><*就*,sufficiency,Back,Intra,81,81,1<不会><有><持久><的><和平>，<两岸><关系><_就,*,82_><难以><改善>。

In the above sentence, the first "就" is matched with the NULL marker, but the second "就" is left as an ADM. This causes an "Other error" and an "ADM/RDM classification error".

The Gross Accuracy (GA) as defined in T'sou et al. (1999) is:

GA = correctly tagged discourse markers / total number of discourse markers = 95.38%

This greatly improves the performance compared with the original GA = 68.89%.

The overgeneration problem (tagged 415, actual 424) is caused by the mismatch of CDMs as RDM pairs, or by the misclassification of CDMs as RDMs. Following are two examples.

<如果,sufficiency,Front,Intra,54,54,1><我们><将,*,55><多元化><理解><成><谁><爱><说><什么><就,*,56><说><什么>，<他人><莫><能干><预>，<那,sufficiency,Back,Intra,57,54,1><就,*,58><好比><主张><在><现代><都市><中><可以><随地><便溺><而,*,59><他人><不得><干涉><一般>。

In this example, "如果" could have matched <将,*,55>, <就,*,56>, or <就,*,58>. Only the <将,*,55> and the <就,*,58> can be eliminated from the candidates according to the "simple rules" mentioned in section 4.1. The system has to choose from <就,*,56> and <那,*,57> to match with "如果". Luckily, the system has given a right choice here.

<一个中国><***是 ,conjunction,Front,Intra,46,46,1***><无可争辩><的><现实>，<NULL,conjunction,Front,Intra,0,49,1><一个中国><原则><***是 ,conjunction,Back,Intra,47,46,1***><包括,*,48><台湾><同胞><在内><的><绝大多数><中国人><的><共同><认识>，<也,conjunction,Back,Intra,49,49,1><得到><世界><上><绝大多数><国家><和><包括,*,50><联合国><在内><的><国际><组织><的><承认>。

The two "是" are misclassified as RDMs, and causes a mismatch of RDM pair. Such errors are difficult to avoid for an automatic system. Without further syntactic/semantic analysis, we can only hope for the ML algorithm to give us a solution from more training data.

6 Conclusion

In order to study discourse markers for use in the automatic summarization of Chinese text, we have designed and implemented the SIFAS system. In this paper, we have focused on the problems of NULL marker location and the classification of RDMs and ADMs. A study on applying machine learning techniques to discourse marker disambiguation is conducted. C4.5 is used to generate decision tree classifiers. Our results indicate that machine learning is an effective approach to improving the accuracy of discourse marker tagging. For interactive use of the system, if we set a threshold for the rule precision and only display those low precision rules for interactive selection, we can greatly speed up the semi-automatic tagging process.

7 References

Chan S., Lai T., Gao W. J. and T'sou B. K. (2000) "Mining Discourse Markers for Chinese Textual Summarization." In Proceedings of the Sixth Applied Natural Language Processing Conference and the North American Chapter of the Association for Computational Linguistics. Workshop on Automatic Summarization, Seattle, Washington, 29 April to 3 May, 2000.

Grosz B.J. and Sidner C. (1986) "Attention, Intention, and the Structure of Discourse," Computational Linguistics 12(3): 175-204.

Hirst G. (1981) "Discourse Oriented Anaphoral Resolution in Natural Language Understanding: A Review." Computational Linguistics 7(2): 85-98.

Hovy E. (1993) "Automated Discourse Generation using Discourse Structure Relations." Artificial Intelligence 63: 341-385.

Hwang C. H. and Schubert L. K. (1992) "Tense Trees as the 'Fine Structure' of Discourse." In Proc. 30th Annual Meeting, Assoc. for Computational Linguistics, pp. 232-240.

Lin H. L., T'sou B. K., H. C. Ho, Lai T., Lun C., C. K. Choi and C.Y. Kit. (1991) "Automatic Chinese Text Generation Based on Inference Trees." In Proc. of ROCLING Computational Linguistic Conference IV, Taipei, pp. 215-236.

Litman D. J. and Allen J. (1990) "Discourse Processing and Commonsense Plans." In Cohen et al.(ed.) Intentions in Communications, pp. 365-388.

Mann W. C. and Thompson S. A (1988) "Rhetorical Structure Theory: Towards a Functional Theory of Text Organization."

Text 8(3): 243-281.

Marcu D. (1997) "From Discourse Structures to Text Summaries." In Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization, Spain, pp. 82-88.

McKeown K. and Radev D. (1995) "Summaries of Multiple News Articles." In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, pp. 74-82.

Ono K., Sumita K. and S. Miike. (1994) "Abstract Generation based on Rhetorical Structure Extraction." In Proceedings of International Conference on Computational Linguistics, Japan, pp. 344-348.

Paice C. D. (1990) "Constructing Literature Abstracts by Computer: Techniques and Prospects." Information Processing and Management 26(1): 171-186.

Quinlan J. Ross (1993) "C4.5 Programs for Machine Learning." San Mateo, CA: Morgan Kaufmann.

T'sou B. K., Ho H. C., Lai B. Y., Lun C. and Lin H. L. (1992) "A Knowledge-based Machine-aided System for Chinese Text Abstraction." In Proceedings of International Conference on Computational Linguistics, France, pp. 1039-1042.

T'sou B. K., Gao W. J., Lin H. L., Lai T. B. Y. and Ho H. C. (1999) "Tagging Discourse Markers: Towards a Corpus based Study of Discourse Marker Usage in Chinese Text" In Proceedings of the 18th International Conference on Computer Processing of Oriental Languages, March 1999, Japan, pp. 391-396.

T'sou B. K., Lin H. L., Ho H. C., Lai T. and Chan T. (1996) "Automated Chinese Full-text Abstraction Based on Rhetorical Structure Analysis." Computer Processing of Oriental Languages 10(2): 225-238.

Tsou, B.K., et al., 1998: 邹嘉彦, 连兴隆, 高维君, 黎邦洋, 何庆昌, "中文篇章中的关联词语及其引导的句子关系的自动标注", ICCIP'98, Beijing, Nov. 18-20, 1998.