Mining Discourse Markers for Chinese Textual Summarization

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

[1 2 4]Language Information Sciences Research Centre
City University of Hong Kong
Tat Chee Avenue, Kowloon Tong,
Hong Kong SAR, China

[3]North Eastern University, China

[1]swkchan@cs.cityu.edu.hk, {[2]cttomlai, [4]rlbtsou}@cpccux0.cityu.edu.hk, [3]wjgao@ramm.neu.edu.cn

Abstract

Discourse markers foreshadow the message thrust of texts and saliently guide their rhetorical structure which are important for content filtering and text abstraction. This paper reports on efforts to automatically identify and classify discourse markers in Chinese texts using heuristic-based and corpus-based data-mining methods, as an integral part of automatic text summarization via rhetorical structure and Discourse Markers. Encouraging results are reported.

1 Introduction

Discourse is understood to refer to any form of language-based communication involving multiple sentences or utterances. The most important forms of discourse of interest to computerized natural language processing are text and dialogue. While discourse such as written text normally appears to be a linear sequence of clauses and sentences, it has long been recognized by linguists that these clauses and sentences tend to cluster together into units, called discourse segments, that are related pragmatically to form a hierarchical structure.

Discourse analysis goes beyond the levels of syntactic and semantic analysis, which typically treats each sentence as an isolated, independent unit. 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. Results of discourse analysis can be used to solve many important NLP problems such as anaphoric reference (Hirst 1981), tense and aspect analysis (Hwang and Schubert 1992), intention recognition (Grosz and Sidner 1986; Litman and Allen 1990), or can be directly applied to computational NLP applications such as text abstraction (Ono et al. 1994; T'sou et al. 1996) and text generation (McKeown 1985; Lin et al. 1991).

Automatic text abstraction has received considerable attention (see Paice (1990) for a comprehensive review). While some statistical approaches have had some success in extracting one or more sentences which can serve as a summary (Brandow et al. 1995; Kupiec et al. 1995; Salton et al. 1997), summarization in general has remained an elusive task. McKeown and Radev (1995) develop a system SUMMONS to summarize full text input using templates produced by the message understanding systems, developed under ARPA human language technology. Unlike previous approaches, their system summarizes a series of news articles on the same event, producing a paragraph consisting of one or more sentences. Endres-Niggemeyer et al. (1995) uses a blackboard system architecture with co-operating object-oriented agents and a dynamic text representation which borrows its conceptual relations from Rhetorical Structure Theory (RST) (Mann and Thompson 1986). Furthermore, connectionist models of discourse summarization have also attracted a lot of attention (Aretoulaki et al. 1998). The main underlying principles are the distributed encoding of concepts and the simulation of human association with a large amount of processing nodes. What is crucial in this approach is to provide a subconceptual layer in the linguistic reasoning.

As in Paice (1990), summarization techniques in text analysis are severely impaired by the absence of a generally accepted discourse model and the use of superstructural schemes is promising for abstracting text. Johnson et al. (1993) describes a text processing system that can identify anaphors so that they may be utilized to enhance sentence selection. It is based on the assumption that sentences which contain non-anaphoric noun phrases and introduce key concepts into the text are worthy of inclusion in an abstract. Ono et al. (1994), T'sou et al. (1992) and Marcu (1997) focus on discourse structure in summarization using the Rhetorical Structure Theory (RST). 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 a RST relation. Summarization is achieved by trimming unimportant sentences on the basis of the relative saliency or rhetorical relations. On the other hand, cohesion can also provide context to aid in the resolution of ambiguity as well as in text summarization (Halliday and Hasan 1976; Morris and Hirst 1991; Hearst 1997). Mani et al. (1998) describes a method based on text coherence which models text in terms of macro-level relations between clauses or sentences to help determine the overall argumentative structure of the text. They examine the extent to which cohesion and coherence can each be used to establish saliency of textual units.

The SIFAS (Syntactic Marker based Full-Text Abstration System) system has been designed and implemented to use discourse markers in the automatic summarization of Chinese. Section 2 provides an introduction to discourse markers in Chinese. An overview of SIFAS is presented in Section 3. In Section 4, we describe a coding scheme for tagging every discourse marker appearing in the SIFAS corpus. In Section 5, we introduce a heuristic-based algorithm for automatic tagging of discourse markers. In Section 6, we describe the application of the C4.5 algorithm to the same task. In Section 7, we present the evaluation results of applying the two algorithms to corpus tagging, followed by a conclusion.

2 Chinese Discourse Markers

Among all kinds of information that may be found in a piece of discourse, discourse markers (also known as discourse connectives, clue words (Reichman 1978; Siegel et al. 1994) or cue phrases (Grosz et al. 1986; Litman 1996) are regarded as the major linguistic device available for a writer to structure a discourse. Discourse markers are expressions which signal a sequential relationship between the current basic message and the previous discourse. Schiffrin (1987) is concerned with elements which mark sequentially dependent units of discourse. She examines discourse markers in interview data, looking specifically at their distribution and their particular interpretation(s). She proposes that these markers typically serve three functions: (i) they index adjacent utterances to the speaker, the hearer, or both; (ii) they index adjacent utterances to prior and/or subsequent discourse; (iii) they work as contextual coordinates for utterances by locating them on one or more planes of her discourse model.

Discourse markers also figure prominently in Chinese which has a tendency to delay topic introduction (Kaplan 1996; Kirkpatrick 1993). Hinds (1982) and Kong (1998) also maintain that the Chinese tendency of delayed topic introduction is heavily influenced by the *qi cheng zhuan he* canonical structure (a Chinese rhetorical pattern). In a study examining rhetorical structure in Chinese, Kirkpatrick (1993) found that several major patterns, favored and considered to be good style by native Chinese writers, are hinted at by Chinese discourse markers. Although the effect of discourse markers in other languages might not be too prominent, there is a great necessity to study discourse markers in Chinese in order to capture the major associated rhetorical patterns in Chinese texts. While the full semantic understanding in Chinese texts is obviously much more difficult to accomplish, the approach using text mining techniques in identifying discourse markers and associated rhetorical structures in a sizeable Chinese corpus will be certainly beneficial to any language processing, such as summarization and knowledge extraction in Chinese.

In Chinese, two distinct classes of discourse markers are useful for identification and interpretation of the discourse structure of a Chinese text: **primary discourse markers** and **secondary discourse markers** (T'sou et al. 1999). Discourse markers can be either words or phrases. Table 1 provides a sample listing of various rhetorical relations and examples considered in this research.

Discourse Markers			
Discourse Type	**Primary Marker**	**Discourse Type**	**Secondary Marker**
Sufficiency	***ruguo*** 'if', ***name*** 'then'	*Summary*	***zong er yan zhi*** 'in one word'
Necessity	***zhiyou*** 'only if', ***cai*** 'only then'	*Contrast*	***shishi shang*** 'in fact'
Causality	***yinwei*** 'because', ***suoyi*** 'therefore'	*Illustration*	***liru*** 'for example'
Deduction	***jiran*** 'given that', ***name*** 'then'	*Specification*	***tebie shi*** 'in particular'
Adversativity	***suiran*** 'although', ***danshi*** 'but'	*Generalization*	***dati er yan*** 'in general'
Concession	***jishi*** 'even if', ***rengran*** 'still'	*Digression*	***wulun ruhe*** 'anyway'
Conjunction	***chule*** 'except', ***jianzhi*** 'also'	*Itemization*	***shouxian*** 'first', ***qici*** 'next'
Disjunction	***huozhe*** 'or', ***huozhe*** 'or'	*Paraphrasing*	***huan ju hua shuo*** 'in other words'
Progression	***budan*** 'not only', ***erqie*** 'but also'	*Equivalence*	***zhengru*** 'just as'
		Enquiry	***nandao*** ('does it mean...')
		Judgment	***kexi*** 'unfortunately'

Table 1 Examples of Discourse Markers and Associated Rhetorical Relations in Chinese

It may be noted that our analysis of Chinese has yielded about 150 discourse markers, and that on the average, argumentative text (e.g. editorials) in Chinese shows more than one third of the discourse segments to contain discourse markers. While primary discourse markers can be paired discontinuous constituents, with each marker attached to one of the two utterances or propositions, the secondary discourse markers tend to be unitary constituents only. In the case of primary discourse markers, it is quite common that one member of the pair is deleted, unless for emphasis. The deletion of both discourse markers is also possible. The recovery process therefore faces considerable challenge even when concerned with the deletion of only one member of the paired discourse markers. Since these discourse markers have no unique lexical realization, there is also the need for disambiguation in a homocode problem.

Moreover, primary discourse markers can also be classified as simple adverbials, as is the case in English:

(1) Even though a child, John is *so* tall that he has problem getting half-fare.

(2) Even though a child, (because) John is tall, *so* he has problem getting half-fare.

In (1), *so* is usually classified as an adverb within a sentence, but in (2) *so* is recognized as marking a change in message thrust at the discourse level.

In the deeper linguistic analysis the two *so*'s may be related, for they refer to a situation involving excessive height with implied consequence which may or may not be stated. In terms of the surface syntactic structure, *so* in (1) can occur in a simple (exclamatory) sentence (e.g. "John is *so* tall!"), but *so* in (2) must occur in the context of complex sentences. Our concern in this project is to identify *so* in the discourse sense as in (2) in contrast to *so* used as an adverb in the sentential sense as in (1). Similar difficulties are found in Chinese, as discussed in Section 7.

3 SIFAS System Architecture

From the perspective of discourse analysis, the study of discourse markers basically involves four distinct but fundamental issues: 1) the occurrence and the frequency of occurrence of discourse markers (Moser and Moore 1995), 2) determining whether a candidate linguistic item is a discourse marker (identification / disambiguation) (Hirschberg and Litman 1993; Siegel and McKeown 1994), 3) determination or selection of the discourse function of an identified discourse marker (Moser and Moore 1995), and 4) the coverage capabilities (in terms of levels of embedding) among rhetorical relations, as well as among individual discourse markers. Discussion of these problems for Chinese compound sentences can be found in Wang et al. (1994).

Previous attempts to address the above problems in Chinese text have usually been based on the investigators' intuition and knowledge, or on a small number of constructed examples. In our current research, we adopt heuristics-based corpus-based approaches, and use machine learning to discover the correlation between various linguistic features and different aspects of discourse marker usage. Our research framework is shown in Figure 1.

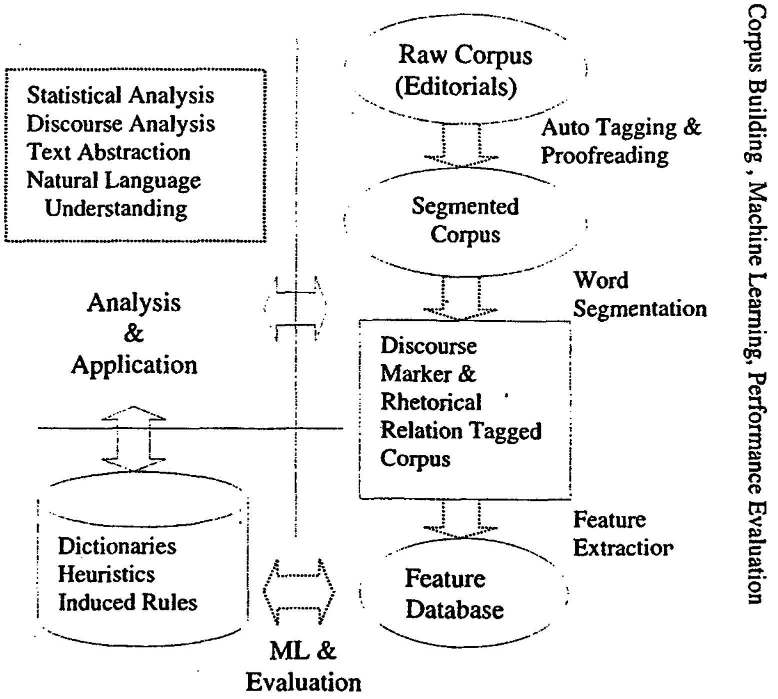

Figure 1 Framework for Corpus-based Study of Discourse Marker Usage in Chinese Text

Data in the segmented corpus are divided into two sets of texts, namely, the training set and the test set, each of which includes 40 editorials in our present research. Texts in the *training set* are manually and semi-automatically tagged to reflect the properties of every Candidate Discourse Marker (CDM). Texts in the *test set* are automatically tagged and proofread. Different algorithms, depending on the features being investigated, are derived to automatically extract the interesting features to form a *feature database*. Machine learning algorithms are then applied to the feature database to generate linguistic rules (decision trees) reflecting the characteristics of various discourse markers and the relevant rhetorical relations. For every induced rule (or a combination of them), its performance is evaluated by tagging the discourse markers appearing in th^ *test set* of the corpus.

4 A Framework for Tagging Discourse Markers

The following coding scheme is designed to encode all and only *Real* Discourse Markers (RDM) appearing in the SIFAS corpus. 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 one of the constituting markers.

RP_i: the *R*elative *P*osition of DM_i. The value of RP_i can be either '*Front*' or '*Back*' denoting the relative position of the marker in the rhetorical relation RR_i.

CT_i: the *C*onnection *T*ype of RR_i. The value of CT_i can be either '*Inter*' or '*Intra*', which indicates that the DM_i functions as a discourse marker in an inter-sentence relation or an Intra-sentence relation.

MN_i: the Discourse *M*arker Sequence *N*umber. The value of MN_i is assigned sequentially from the beginning of the processed text to the end.

RN_i: the Rhetorical *R*elation Sequence *N*umber. The value of RN_i is assigned sequentially to the corresponding rhetorical relation RR_i in the text.

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 (ADM) that do not function as real discourse markers in a text, a different 3-tuple coding scheme is used to encode them:

ADM_i = < LI_i, *, SN_i > where,

LI_i: the *L*exical *I*tem of the ADM.

SN_i: the *S*equence *N*umber of the ADM.

To illustrate the above coding scheme, consider the following examples of encoded sentences where every CDM has been tagged to be either a 7-tuple or a 3-tuple.

Example 1

<***youyu*** ('because'),*Causality, Front, Intra, 2, 2, 1*> Zhu Pei ('Jospin') zhengfu ('government') taidu ('attitude') qiangying ('adamant'), chaoye ('government-public') duikang ('confrontation') yue-yan-yue ('more-develop-more') -lie ('strong'), <*NULL, Causality, Back, Intra, 0, 2, 1*> gongchao ('labour unrest') <*yi* ('with'), *, 1> liaoyuan ('burn-plain') zhi ('gen') shi 'tendency' xunshu 'quick' poji 'spread to' ge ('every') hang ('profession') ge ('every') ye ('trade').

'As a result of the adamant attitude of the Jospin administration, confrontation between the government and the public is becoming worse and worse. Labour unrest has spread quickly to all industrial sectors.'

From the above tagging, we can immediately obtain the discourse structure that the two clauses encapsulated by the two discourse markers *youyu* (with sequence number 2) and *NULL* (with sequence number 0). They have formed a causality relation (with sequence number 2). We denote this as a binary relation

***Causality*(FrontClause(2), BackClause(2))**

where **FrontClause(*n*)** denotes the discourse segment that is encapsulated by the **Front** discourse marker of the corresponding rhetorical relation whose sequence number is n. **BackClause(*n*)** can be defined similarly. Note that although *yi* is a CDM, it does not function as a discourse indicator in this sentence. Therefore, it is encoded as an apparent discourse marker.

Example 2

<***dan*** ('however'), *Adversativity, Back, Inter, 17, 14, 1*><***ruguo*** 'if', *Sufficiency, Front, Inter, 18, 15, 1*> Zhu Pei ('Jospin') zhengfu ('government') cici ('this time') zai ('at') gongchao ('labour unrest') mianqian ('in the face of') tuique ('back down'), <*NULL, Sufficiency, Back, Inter, 0, 15, 1*> houguo ('result') <***geng***,('more'), *, 3> shi bukan ('is unbearable') shexian ('imagine').

'However, if the Jospin administration backs down in the face of the labour unrest, the result will be terrible.'

From the above tagging, we can obtain the following discourse structure with embedding relations:

***Adversativity*(&F(14),**

***Sufficiency*(FrontClause(15), BackClause(15)))**

where **&F(*n*)** denotes the **Front** discourse segment of an inter-sentence rhetorical relation whose sequence number is ***n***. We can define **&B(*n*)** similarly.

5 Heuristic-based Tagging of Discourse Markers

In the previous section, we have introduced a coding scheme for CDMs, and have explained how to automatically derive the discourse structure from sentences with tagged discourse markers. Now, the problem we have to resolve is: Is there an algorithm that will tag the markers according to the above encoding scheme?

To derive such an algorithm, even an imperfect one, it is necessary that we have knowledge of the usage patterns and statistics of discourse markers in unrestricted texts. This is exactly what project SIFAS intends to achieve as explained in Section 3. Instead of completely relying on a *human encoder* to encode all the training texts in the SIFAS corpus, we have experimented with a simple algorithm using a small number of heuristic rules to automatically encode the CDMs. The algorithm is a straightforward matching algorithm for rhetorical relations based recognition of their constituent discourse markers as specified in the Rhetorical Relation Dictionary (T'sou et al. 1999). The following principles are adopted by the heuristic-based algorithm to resolve ambiguous situations encountered in the process of matching discourse markers:

(1) ***Principle of Greediness***: When matching a pair of CDMs for a rhetorical relation, priority is given to the first matched relation from the left.
(2) ***Principle of Locality***: When matching a pair of CDMs for a rhetorical relation, priority is given to the relation where the distance between its constituent CDMs is shortest.
(3) ***Principle of Explicitness***: When matching a pair of CDMs for a rhetorical relation, priority is given to the relation that has both CDMs explicitly present.
(4) ***Principle of Superiority***: When matching a pair of CDMs for a rhetorical relation, priority is given to the inter-sentence relation whose back discourse marker matches the first CDM of a sentence.
(5) ***Principle of Back-Marker Preference***: this principle is applicable only to rhetorical relations where either the front or the back marker is absent. In such cases, priority is given to the relation with the back marker present.

Application of the above principles to process a text is in the order shown, with the exception that the principle of greediness is applied whenever none of the other principles can be used to resolve an ambiguous situation. The following pseudo code realizes principles 1, 2 and 3:

```
I := 1 ;
while I < NumberOfCDMsInTheSentence do
begin
for J:=1 to NumberOfCDMsInTheSentencen -
I do
   if ((not CDMs[J].Tagged) and (not
      CDMs[J+I].Tagged)) then
      Matching(CDMs[J], CDMs[J+I]) ;
   I := I + 1 ;
 end ;
```

The following code realizes principles 1, 4 and 5:

```
for I:=1 to NumberOfCDMsInTheSentence do
begin
  if (not CDMs[I].Tagged) then
     Matching(NULL, CDMs[I]) ;
  if (not CDMs[I].Tagged) then
     Matching(CDMs[I], NULL) ;
end ;
```

In the above pseudo codes, **CDMs[]** denotes the array holding the candidate discourse markers, and the Boolean variable **Tagged** is used to indicate whether a CDM has been tagged. Furthermore, the procedure **Matching()** is to examine whether the first word or phrase appearing in a sentence is an inter-sentence **CDMs[I]**.

6 Mining Discourse Marker Using Machine Learning

Data mining techniques constitute a field dedicated to the development of computational methods underlying learning processes and they have been applied in various disciplines in text processing, such as finding associations in a collection of texts (Feldman and Hirsh 1997) and mining online text (Knight 1999). In this section, we focus on the problem of discourse marker disambiguation using *decision trees* obtained by machine learning techniques. Our novel approach in mining Chinese discourse markers attempts to apply the C4.5 learning algorithm, as introduced by Quinlan (1993), in the context of non-tabular, unstructured data. A decision tree consists of nodes and branches connecting the nodes. The nodes located at the bottom of the tree are called leaves, and indicate classes. The top node in the tree is called the root, and contains all the training examples that are to be divided into classes. In order to minimize the branches in the tree, the best attribute is selected and used in the test at the root node of the tree. A descendant of the root node is then created for each possible value of this attribute, and the training examples are sorted to the appropriate descendant node. The entire process is then repeated using the training examples associated with each descendant node to select the best attribute for testing at that point in the tree. A statistical property, called *information gain*, is used to measure how well a given attribute differentiates the training examples according to their target classificatory scheme and to select the most suitable candidate attribute at each step while expanding the tree.

The attributes we use in this research 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. All these five attributes are discrete. The following are two examples:

- ",", dan 'but', **youyu** 'since', Xianggang 'Hong Kong', de 'of', T.
- zhe 'this', yi 'also', **zhishi** 'is only', Xianggang 'Hong Kong', de 'of', F.

where "T" denotes the CDM **youyu** as a discourse marker in the given context, and "F" denotes that **zhishi** is not a discourse marker.

In building up a decision tree in our application of C4.5 to the mining of discourse markers, entropy, first of all, is used to measure the homogeneity of the examples. For any possible candidate *A* chosen as an attribute in classifying the training data *S*, *Gain*(*S*, *A*) information gain, relative to a data set *S* is defined. This information gain measures the expected reduction in entropy and defines one branch for the possible subset S_i of the training examples. For each subset S_i, a new test is then chosen for any further split. If S_i satisfies a stopping criterion, such as all the element in S_i belong to one class, the decision tree is formed with all the leaf nodes associated with the most frequent class in *S*. C4.5 uses *arg max*(*Gain*(*S*, *A*)) or *arg max*(*Gain Ratio*(*S*, *A*)) as defined in the following to construct the minimal decision tree.

$$Entropy(S) = -\sum_{i=1}^{c} - p_i \log_2 p_i \qquad \text{(Eqn. 1)}$$

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \qquad \text{(Eqn. 2)}$$

$$Gain\ Ratio = \frac{Gain(S,A)}{SplitInformation(S,A)} \qquad \text{(Eqn. 3)}$$

where $SplitInformation = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$, S_i is subset of *S* for which *A* has value v_l

In our text mining, according to the number of times a CDM occurs in the 80 tagged editorials, we select 75 CDMs with more than 10 occurrences. To avoid decision trees being over-fitted or trivial, for **F2**, **F1**, **B1** and **B2**, only values of attributes with frequency more than 15 in the corpus are used in building the decision trees. We denote all values of attributes with frequency less than 15 as '**Other**'. If a CDM is the first, the second or the last word of a sentence, values of **F2**, **F1**, or **B2** will be null, we denote a null-value as "*". The following are two other examples:

- "*", "*", **zheyang** 'thus', ",", Other, T.
- "*", "*", **zheyang** 'thus', Other, de 'of', F.

7 Evaluation

7.1 Evaluation of Heuristic-based Algorithm

In order to evaluate the effectiveness of the heuristic-based algorithm, we randomly selected 40 editorials from Ming Pao, a Chinese newspaper of Hong Kong, to form our test data. Only editorials are chosen because they are mainly argumentative texts and their lengths are relatively uniform.

The steps of evaluation consist of: 1) tagging all of the test data using the heuristic-based algorithm, and 2) proofreading, correcting and recording all the tagging errors by a human encoder. The resulting statistics include, for each editorial in the test data, the number of lexical items (#LItms), the number of sentences (#Sens), the number of discourse markers (#Mrkrs), and the number of sentences containing at least one discourse marker (#CSens). Table 2 shows the minimum, maximum and average values of these characteristics. The ratio of the average number of discourse markers to the average number of lexical items is **4.37%**, and the ratio of the average number of sentences containing at least one discourse marker to the average number of sentences is **62.66%**.

	#LItms	#Mrkrs	#Sens	#CSens
MIN	466	14	11	6
MAX	1082	52	45	26
AVERAGE	676.25	29.58	22.15	13.88

Table 2 Characteristics of the Test Data

Our evaluation is based on counting the number of discourse markers that are correctly tagged. For incorrectly tagged discourse markers, we classify them according to the types of errors that we have introduced in T'sou et al. (1999). We define two evaluation metrics as follows: ***Gross Accuracy*** (GA) is defined to be the percentage of correctly tagged discourse markers to the total number of discourse markers while ***Relation-Matching Accuracy*** (RMA) is defined to be the percentage of correctly tagged discourse markers to the total number of discourse markers minus those errors caused by non-markers and unrecorded markers. The results for our testing data have GA = 68.89% and RMA = 95.07%.

Since the heuristic-based algorithm does not assume any knowledge of the statistics and behavioral patterns of discourse markers, our GA demonstrates the usefulness of the algorithm in alleviating the burden of human encoders in developing a sufficiently large corpus for the purpose of studying the usage of discourse markers.

In our experiment, most errors come from tagging non-discourse markers as discourse markers (T'sou et al. 1999). This is due to the fact that, similar to the question of cue phrase polysemy (Hirschberg and Litman 1993), many Chinese discourse markers have both discourse senses and alternate sentential senses in different utterances. For example:

- ... Zhe ('this') **buguo** shi ('only is') yi ('one') ge ('classifier') wanxiao ('joke')
...('This is only a joke'.) (sentential sense)
- ...**Buguo** ('however'), wo ('I') bu ('neg') zheyang ('thus') renwei ('consider')
...('But I don't think so.') (discourse sense)

7.2 Evaluation of Decision Tree Algorithm (with C4.5)

In Section 6, we discuss how machine learning techniques have been applied to the problem of discourse marker disambiguation in Chinese.

In our experiment, there are a total of 2627 cases. In our decision tree construction, we use 75 percent of the total cases as a training set, and the remaining 25 percent of cases as a test set. Many decision trees can be generated by adjusting the parameters in the learning algorithm. Many decision trees generated in our experiment have an accuracy around 80% for both the training set and the test set. Figure 2 shows one of the possible decision trees in our experiment. The last branch of the decision tree

```
F1 = danshi 'but'
| CDM in {ru 'if', reng 'still', geng 'even more', que
        'however'}: F (6/0)
| CDM in {chule 'except', youyu 'since', ruo 'if'} : T
   (4/0)
```

can be explained as:

```
if (F1 = danshi 'but') then
  if (CDM in {ru 'if', reng 'still', geng 'even more', que
   'however'}) then classify as F
else
  if (CDM in {chule 'except', youyu 'since', ruo 'if'})
  then classify as T
```

```
Decision Tree: (Size = 38, Items = 1971, Errors = 282)
F1 in {di, ye, yi} :   F (25/5)
F1 in {, shi, ;}   :   T (712/131)
F1 = Other:
.
.
.
F1 = danshi :
| CDM in {ru, reng, geng, que}    : F (7/10)
| CDM in {chule, youyu, ruo}  :   T (4/0)

Evaluation on training data from Data. Data (1971 cases):
    Classified results:
     T       F     | <- Classified
    937     125    | Class : T
    157     752    | Class : F     Errors : 282 (14.3%)

Evaluation on testing data from Data. Test (656 cases):
    Classified results:
    T       F      |<- Classified
    293     62     |Class : T
    68      233    |Class : F      Errors : 130 (19.8%)
```

Figure 2 An Example of Decision Trees

The two numbers in the brackets denote the number of cases covered by the branch and the number of cases being misclassified respectively. The results of our experiment will be elaborated on in future, when we shall also explore the application of machine learning techniques to recognizing rhetorical relations on the basis of discourse markers, and extracting important sentences from Chinese text.

8. Conclusion

We discuss in this paper the use of discourse markers in Chinese text summarization. Discourse structure trees with nodes representing RST (Rhetorical Structure Theory) relations are built and summarization is achieved by trimming unimportant sentences on the basis of the relative saliency or rhetorical relations. In order to study discourse markers for use in the automatic summarization of Chinese, we have designed and implemented the SIFAS system. We investigate the relationships between various linguistic features and different aspects of discourse marker usage on naturally occurring text. An encoding scheme that captures the essential features of discourse marker usage is introduced. A heuristic-based algorithm for automatic tagging of discourse markers is designed to alleviate the burden of a human encoder in developing a large corpus of encoded texts and to discover potential problems in automatic discourse marker tagging. A study on applying machine learning techniques to discourse marker disambiguation is also conducted. C4.5 is used to generate decision tree classifiers. Our results indicate that machine learning is a promising approach to improving the accuracy of discourse marker tagging.

9 Acknowledgement

Support for the research reported here is provided through the Research Grants Council of Hong Kong under Competitive Earmarked Research Grants (CERG) No. 9040067, 9040233 and 9040326.

10 References

Aretoulaki M., Scheler G. and Brauer W. (1998) "Connectionist Modeling of Human Event Memorization Processes with Application to Automatic Text Summarization." In *Proceedings of AAAI Spring Symposium on Intelligent Text Summarization*, Stanford, pp. 148-150.

Brandow R., Mitze K. and Rau L. F. (1995) "Automatic Condensation of Electronic Publications by Sentence Selection." *Information Processing and Management*, 31(5): 675-685.

Endres-Niggemeyer B., Maier E. and Sigel A. (1995) "How to Implement a Naturalistic Model of Abstracting: Four Core Working Steps of an Expert Abstractor." *Information Processing and Management*, 31(5): 631-674.

Feldman R. and Hirsh H. (1997). "Finding associations in collections of text." In R.S. Michalski I. Bratko and Kubat M. (Eds.), *Machine Learning and Data Mining: Methods and Applications*, pp. 224-240. Wiley.

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

Halliday M. A. K. and Hasan R. (1976) *Cohesion in English*, Longman.

Hearst M. A. (1997) "Texttiling: Segmenting Text into Multi-paragraph Subtopic Passages." *Computational Linguistics*, 23(1):33-64.

Hinds J. (1982) "Inductive, deductive, quasi-inductive: Expository writing in Japanese, Korean, Chinese, and Thai." In U. Connor and A.M. Johns (Eds.). *Coherence in Writing*, pp. 89-109. TESOL publisher.

Hirschberg J. and Litman D. (1993) "Empirical Studies on the Disambiguation of Cue Phrases." *Computational Linguistics* 19(3): 501-530.

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.

Johnson F. C., Paice C. D., Black W. J. and Neal A. P. (1993) "The Application of Linguistic Processing to Automatic Abstract Generation." *Journal of Document and Text Management* 1: 215-241.

Kaplan R. B. (1996) "Cultural though patterns in intercultural education." *Language Learning*, 1&2: 1-20.

Kirkpatrick A. (1993) "Information sequencing in modern standard Chinese in a genre of extended spoken discourse." *Text* 13(3): 423-453.

Kong K.C.C. (1998) "Are simple business request letters really simple? A comparison of Chinese and English business request letters." *Text* 18(1): 103-141.

Knight K. (1999) "Mining online text." *Communications of the ACM* 42(11): 58-61.

Kupiec J., Pedersen J., and Chen F. (1995) "A Trainable Document Summarizer." In

Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, pp. 68-73.

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.

Litman D. J. (1996) "Cue Phrase Classification Using Machine Learning." *Journal of Artificial Intelligence Research* 5: 53-94.

Mani I., Bloedorn E. and B. Gates (1998) "Using Cohesion and Coherence Models for Text Summarization." In *Proceedings of AAAI Spring Symposium on Intelligent Text Summarization*, Stanford, pp. 69-76.

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.

McKeown K. R. (1985) "Discourse Strategies for Generating Natural-Language Text." *Artificial Intelligence* 27(1): 1-41.

Morris J. and Hirst G. (1991) "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text." *Computational Linguistics* 17(1): 21-48.

Moser M. and Moore J. D. (1995) "Investigating Cue Selection and Placement in Tutorial Discourse." In *Proceedings of ACL'95*, pp. 130-135.

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.

Reichman R. (1978) "Conversational Coherence." *Cognitive Science* 2(4): 283-328.

Salton G., Singhal A., Mitra M. and Buckley C. (1997) "Automatic Text Structuring and Summarization." *Information Processing and Management* 33(2): 193-207.

Schiffrin D. (1987) *Discourse Markers.* Cambridge: Cambridge University Press.

Siegel E. V. and McKeown K. R. (1994) "Emergent Linguistic Rules from Inducing Decision Trees: Disambiguating Discourse Clue Words." In *Proceedings of AAAI*, pp. 820-826.

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.

Wang W. X., Zhang X. C., Lu M. Y. and Cheng H. Y. (1994) "Xian Dai Han Yu Fu Ju Xian Jie (A New Analysis of Complex Sentences in Modern Standard Chinese)", Hua Dong Shi Fan Da Xue Chu Ban She, 1994.