Analyses of the Association between Discourse Relation and Sentiment Polarity with a Chinese Human-Annotated Corpus

Hen-Hsen Huang Chi-Hsin Yu Tai-Wei Chang Cong-Kai Lin Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan
{hhhuang, jsyu, twchang, cklin}@nlg.csie.ntu.edu.tw;
hhchen@ntu.edu.tw

Abstract

Discourse relation may entail sentiment information. In this work, we annotate both discourse relation and sentiment information on a moderate-sized Chinese corpus extracted from the ClueWeb09. Based on the annotation, we investigate the association between the relation type and the sentiment polarity in Chinese and interpret the data from various aspects. Finally, we highlight some language phenomena and give some remarks.

1 Introduction

A discourse relation indicates how two arguments (i.e., elementary discourse units) cohere to each other. Various discourse relations were defined according to different taxonomy (Carlson and Marcu, 2001; Carlson et al., 2002; Prasad et al., 2008). In the work of the Penn Discourse Treebank 2.0 annotation, Prasad et al. (2008) labeled four grammatical classes of connectives in English, including subordinating conjunctions, coordinating conjunctions, adverbial connectives, and implicit connectives. Besides, the sense of each connective was also tagged. They defined three levels of sense hierarchy for the connectives. The four classes on the top level are *Temporal*, *Contingency*, *Comparison*, and *Expansion*.

There are *explicit* and *implicit* uses of discourse relations. An explicit discourse relation indicates the arguments are connected with an overt discourse marker (i.e., connective). A connective joins two discourse units such as phrases, clauses, or sentences together. For example, the word *however* is a common connective that indicates a *Comparison* relation between two arguments. The sense of a discourse marker denotes how its two arguments cohere. In other words, a discourse marker presents the relation of its two arguments.

In other cases, discourse marker is absent from an implicit relation. However, readers can still infer the relation from its argument pair. To resolve implicit discourse relations, i.e., without the information from discourse markers, is more challenging (Lin et al., 2009; Zhou et al., 2010).

Hutchinson (2004) pointed out the properties of a discourse marker from three dimensions, including polarity, veridicality, and type. The polarity of a discourse marker indicates the sentiment transition of its two arguments. Veridicality, the second dimension of a discourse marker, specifies whether both the two arguments are true or not. Type, similar to the sense which is annotated in the PDTB, is the third dimension of a discourse marker.

Our previous work (Huang and Chen, 2012a; Huang and Chen, 2012b) addressed the interaction between the sentiment polarity and the discourse structure in Chinese. Consider (S1), which consists of three clauses and forms a nested discourse structure shown in Figure 1.

(S1) 管理處雖然嘗試要讓長期以來作為大台北後花園的陽明山區更回歸自然 (Although the management office tried to make the Yangmingshan area a more natural environment as the long-term garden of Taipei)，但隨著週休二日、經濟環境改善 (but due to the two-day weekend and the improved economic conditions)，遊客帶來停車、垃圾等間接影響卻更嚴重 (the issues of tourist parking, garbage, and other indirect effects become more serious)。

The second and the third clauses form a *Contingency* relation with a sentiment polarity transition from Positive to Negative. Furthermore,

Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, pages 70–78,
Sofia, Bulgaria, August 8-9, 2013. ©2013 Association for Computational Linguistics

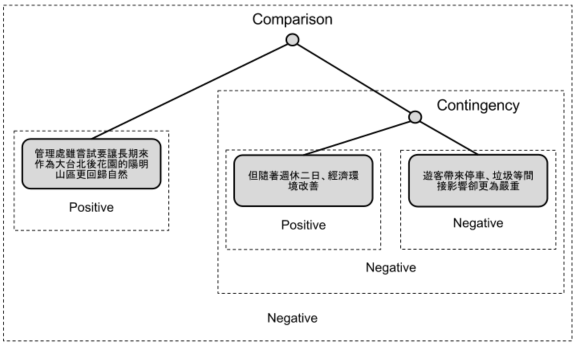
Comparison
Contingency
管理處雖嘗試要讓長期來作為大台北後花園的陽明山區更回歸自然
Positive
但隨著週休二日、經濟環境改善
Positive
遊客帶來停車、垃圾等間接影響卻更為嚴重
Negative
Negative
Negative

Figure 1: Discourse structure and sentiment polarities of (S1).

these two clauses also constitute one of the arguments of a *Positive-Negative Comparison* relation. As the PDTB 2.0 annotation manual suggests (Prasad, et al., 2007), a *Comparison* relation is established to emphasize the differences between two arguments. Therefore, it is expected that the two arguments of a *Comparison* relation are relatively likely to have the opposing polarity states (i.e., *Positive-Negative* or *Negative-Positive*). On the other hand, the two arguments of an *Expansion* relation are relatively likely to belong to the same polarity states (e.g., *Positive-Positive* or *Neutral-Neutral*).

Discourse relation recognition (Hernault et al., 2010; Soricut and Marcu, 2003) and sentiment analysis (Pang and Lee, 2008) have attracted much attention recently. Due to the limitation of the resources, the research on Chinese discourse relation analysis is relatively rare. In our previous work, we annotated a collection of Chinese discourse corpora, namely NTU Chinese Discourse Resources (http://nlg.csie.ntu.edu.tw/ntu-discourse/), for inter-sentential and intra-sentential discourse relation recognition (Huang and Chen, 2011; Huang and Chen, 2012a). However, no sentiment information is labeled in these corpora. In another work (Huang and Chen, 2012b), we proposed an annotation scheme to construct a Chinese discourse corpus with rich information including sentiment polarities, but the corpus is still under construction due to its complexity. Zhou and Xue (2012) did PDTB-style Chinese discourse corpus annotation, but the corpus is also not available yet.

In this paper, we annotate a moderate-sized Chinese corpus with the information of discourse relations and sentiment polarities. Total 7,638 sentences are sampled from the ClueWeb09. We review the results of annotation and analyze some language phenomena found in the corpus.

The rest of this paper is organized as follows. In Section 2, we introduce the ClueWeb corpus and a dictionary of Chinese discourse markers. In Section 3, the criteria to sample instances and the annotation scheme are shown. We analyze the language phenomena found in the annotated data and discuss the correlation between discourse relations and sentiment polarities in Section 4. Finally, we conclude the remarks in Section 5.

2 Linguistic Resources

The PDTB is a popular dataset used in the English discourse research. In contrast, no Chinese discourse corpus is publicly available at present. To construct a Chinese discourse corpus, we sample instances from a huge Chinese corpus (Yu et al., 2012). This corpus was developed based on the ClueWeb09 dataset, where Chinese material is the second largest. It contains a total of 9,598,430,559 POS-tagged sentences in 172,298,866 documents.

In this paper, only the explicit discourse relations are concerned. A dictionary of discourse markers is consulted to extract the instances of explicit discourse relations from the ClueWeb. This Chinese discourse marker dictionary is developed based on Cheng and Tian (1989), Cheng (2006) and Lu (2007). Table 1 shows an overview of the discourse marker dictionary. It contains 808 words and word pairs mapped into the PDTB four top-level classes (Cheng and Tian, 1989; Wolf and Gibson, 2005). Besides the types of discourse relations, we further classify the markers into three groups of scopes shown in the second column, including *Single word*, *Intra-sentential*, and *Inter-sentential*, according to their grammatical usages. The *Single word* group contains those individual words used as discourse markers. The *Intra-sentential* group contains pairs of words that occur inside the same sentence and denote a discourse relation. Here, a Chinese sentence is defined as a sequence of successive words that is ended by a period, a question mark, or an exclamation mark. The clauses of a sentence are delimited by commas. The *Inter-sentential* discourse markers are similar to the *Intra-sentential* ones, but the two words of a pair individually appear in different sentences. Some discourse markers can be used as both Inter-sentential and Intra-sentential. In this work, the Inter-sentential only discourse markers are excluded because we only concern the discourse relation occurring within a sentence. The third column lists the number of discourse markers for each scope under each PDTB class, and the fourth column gives some examples.

PDTB Class	Scope	# Markers	Examples
Expansion	Single word	177	另外 (besides), 抑或 (or), 不只 (not only), 例如 (such as)
	Intra-sentential	106	一方面…一方面 (on the one hand ... on the other hand), 不是…而是 (not ... but), 不只…也 (not only ... also)
	Inter-sentential	26	首先…再者 (first ... second), 或…或許 (or ... perhaps), 不只…不只 (not only ... not only)
Temporal	Single word	41	接著 (then)
	Intra-sentential	80	最初…最後 (first ... finally)
	Inter-sentential	30	最初…現在 (first ... now)
Comparison	Single word	34	即使 (even if)
	Intra-sentential	38	儘管…但 (although ... but)
	Inter-sentential	15	雖說…其實 (in spite of ... in fact)
Contingency	Single word	67	因為 (because), 如 (if), 假設 (suppose), 以免 (in order to avoid)
	Intra-sentential	180	因…而 (because ... then), 如…則 (if then), 凡…可 (any ... can)
	Inter-sentential	14	既然…於是 (since ... then), 至少…不然 (at least ... otherwise)

Table 1: Overview of a Chinese discourse marker dictionary.

3 Annotation

Based on the Chinese part of the ClueWeb09 (Yu et al., 2012), we sample a moderate-sized data with some criteria and annotate them with the information of discourse relations and sentiment polarities.

3.1 Sampling a reliable dataset

Discourse relations may be explicit or implicit, and a sentence may contain more than one discourse marker. Multiple discourse relations occurring in a sentence will make the annotation more complex. In this work, we focus on the correlation between discourse relations and sentiment polarity. To get a reliable dataset for analysis, we sample sentences based on the following three criteria.

1. A sentence should contain only two clauses.
2. A sentence should contain exact one discourse marker shown in the Chinese discourse marker dictionary. We match the discourse marker on the word level. For the *Single word* markers, the marker can appear in either of the clauses. For the pairwise markers, the first word should appear in the first clause, and the second word should appear in the second one.
3. The lengths of both clauses in a sentence are no more than 20 Chinese characters.

As shown in Figure 1, the sentiment polarity determination is more challenging when more than one discourse relation is involved in a sentence. In order to facilitate the analysis, we focus on those sentences that contain exact one discourse marker. The limitation of clause length is also applied to avoid the noise from implicit discourse relation. Based on a preliminary statistics, we find that most clauses in the Chinese part of the ClueWeb (Yu et al., 2012) are no longer than 20 Chinese characters shown in Figure 2.

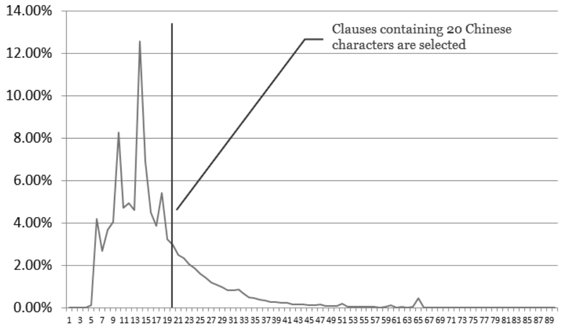

Figure 2: Length distribution in the ClueWeb.

3.2 Annotation scheme

Using the criteria described in Section 3.1, total 7,638 instances are randomly selected from the ClueWeb, and 87 native speakers annotate these instances. Each instance is shown to three annotators. The annotator labels the polarities of the first clause, the second clause, and the whole instance with *Negative*, *Neutral*, and *Positive*. In addition, the discourse relation between the two clauses is also labeled with *Temporal*, *Contingency*, *Comparison*, and *Expansion*. For each target sentence, the annotation is based on the information from the sentence only. The sentences are not given to annotators. Finally, the majority of each label is taken. For example, the polarity p_1 of the first clause in the instance (S2) is labeled as *Positive*, the polarity p_2 of the second clause is labeled as *Negative*, the resulting polarity p_w of the whole sentence is also labeled as *Negative*, and the discourse relation between the two clauses is labeled as *Comparison*.

(S2) 法國品牌的汽車在本土市場的佔有率雖然過半 (Although French brand cars share more than half of the domestic market share)，但市場份額持續萎縮 (but the market share continued to shrink)。

The inter-agreements of p_1, p_2, p_w, and discourse relation among annotators are 0.49, 0.50, 0.47, and 0.41 in Fleiss' Kappa values, respectively (all are moderate agreement). The resulting corpus is publicly available on the website of NTU Chinese Discourse Resources[1].

4 Results and Discussion

To investigate the corpus annotated with discourse relation and sentiment polarity, we firstly give an overview of results with respect to these two types of linguistic phenomena. And then, the most frequent discourse markers for each class of discourse relations are discussed. Finally, we reorganize the results to several aspects and discuss the association between discourse relations and sentiment polarities.

4.1 Overview of the annotated corpus

The distribution of the discourse relations versus the polarities of whole sentence (p_w) is shown in Table 2. Compared to the distributions of discourse relations in the Penn Discourse Treebank (Prasad et al., 2008) shown in Table 3, the explicit Chinese discourse corpus is more similar to the whole English corpus. The instances of *Expansion* form the largest set among four discourse relation classes. In Chinese, the instances of *Expansion* are even more. *Temporal* is the most infrequent relation which has close frequencies in both corpora. The different characteristic is the frequency of *Comparison* relation. In our Chinese corpus, the frequency of *Comparison* relation is about half of that in the PDTB.

In Table 2, the symbol † is used to highlight the relatively major polarity of each relation. The symbol ‡ is marked when the polarity is the majority (i.e., with a frequency greater than 50%). Near half (49.11%) of the instances belong to *Neutral*. Neutral statements are major in *Temporal* and *Expansion* classes. On the other hand, *Comparison* is the relation which is most involved in expressing sentiment, negative sentiment in particular. *Contingency* is second to *Comparison* in expressing sentiment.

The distribution of the discourse relations versus (p_1, p_2), the sentiment polarity transitions between two clauses, is shown in Table 4. *Neutral-Neutral* is the most frequent polarity transition in all relations. More than half of the *Temporal* instances are *Neutral-Neutral*. The reason may be that the *Temporal* relations are usually used in the sentences that describe the objective facts of the past, present, or the future. In such sentences, the sentiments are relatively rare. On the other hand, the sentences of *Comparison* and *Contingency* occur more in the critical and analytical scenarios.

Although the most frequent transition of *Comparison* is also *Neutral-Neutral* (23.14%), the other three types of transitions, *Positive-Negative*, *Neutral-Negative*, and *Negative-Positive*, have close frequencies of 22.71%, 16.90%, and 15.72%, respectively. Moreover, *Negative* polarity is involved in all these three transitions in one of their clauses.

The relations between p_1, p_2, and p_w are also interesting. Table 5 shows the top 10 most frequent correlations of the polarities (p_1, p_2, p_w) of the first clause, the second clause, and the whole sentence. On the one hand, it is not surprising that most instances belong to (*Neutral*, *Neutral*, *Neutral*). On the other hand, it is worthy of noting that p_2 and p_w are identical in the top eight types of combinations in Table 5. In other words, the resulting sentiment polarity of a two-clause sentence is mostly consistent with the polarity of

Relation	#	%	Neu (%)	Pos (%)	Neg (%)
Temporal	849	11.12	‡60.66	22.38	16.96
Contingency	1,598	20.92	†44.74	26.97	28.29
Comparison	929	12.16	33.37	27.88	†38.75
Expansion	4,262	55.80	‡51.88	31.75	16.38
Overall	7,638	100.00	†49.11	29.24	21.65

Table 2: Distribution of discourse relations vs. polarities of whole sentences.

Relation	Only Explicit Cases		Total	
	#	%	#	%
Temporal	3,612	18.88	4,650	12.71
Contingency	3,581	18.72	8,042	21.98
Comparison	5,516	28.83	8,394	22.94
Expansion	6,424	33.58	15,506	42.38
Overall	19,133	100.00	36,592	100.00

Table 3: Distribution of discourse relations in the Penn Discourse TreeBank 2.0.

[1] http://nlg.csie.ntu.edu.tw/ntu-discourse/

PDTB Class	#	Distribution of each type of sentiment polarity transition (p_1, p_2) (%)								
		Neu Neu	Pos Neu	Neg Neu	Neu Pos	Pos Pos	Neg Pos	Neu Neg	Pos Neg	Neg Neg
Temporal	849	‡57.01	1.53	2.12	16.37	3.53	2.36	12.72	1.06	3.30
Contingency	1,598	†35.42	3.69	5.88	13.70	10.45	2.32	11.64	1.81	15.08
Comparison	929	†23.14	2.69	2.48	8.61	3.12	15.72	16.90	22.71	4.63
Expansion	4,262	†48.33	2.86	1.92	14.24	16.19	0.59	7.86	0.63	7.37
Overall	7,638	†43.53	2.87	2.84	13.68	11.99	2.99	10.29	3.61	8.20

Table 4: Distribution of discourse relations vs. types of sentiment transitions.

p_1	p_2	p_w	Occurrences
Neutral	Neutral	Neutral	3,268
Neutral	Positive	Positive	945
Positive	Positive	Positive	908
Neutral	Negative	Negative	706
Negative	Negative	Negative	614
Positive	Negative	Negative	204
Negative	Positive	Positive	199
Negative	Neutral	Neutral	125
Positive	Neutral	Positive	121
Neutral	Positive	Neutral	99

Table 5: Most frequent (p_1, p_2, p_w) combinations.

	$p_1 = p_w$	$p_1 \neq p_w$	Total
$p_2 = p_w$	62.71%	29.79%	92.50%
$p_2 \neq p_w$	5.51%	1.99%	7.50%
Total	68.22%	31.78%	100.00%

Table 6: Correlations between (p_1,p_w) and (p_2,p_w).

the second clause. Table 6 shows the correlations of sentiment polarities between clauses and the whole sentence. Total 92.50% of instances belong to the case ($p_2 = p_w$), where the polarity of the second clause is identical to the polarity of the whole sentence. In Chinese writing, putting the important part of a sentence at the end of the sentence is very common.

4.2 Frequent discourse markers

The top discourse markers in our Chinese corpus are shown in Table 7. For each PDTB class, the five most frequent discourse markers are listed. In each row of the table, its number of occurrences and the distribution of its nine sentiment polarity transitions are given. Note that there are three polarities, i.e., *positive*, *neutral*, and *negative*. The relatively major sentiment polarity transition of each discourser maker is labeled with the symbol †. The symbol ‡ is marked when the sentiment polarity is the majority, i.e., its ratio is greater than 50%.

Some discourse markers are the top markers in more than one discourse relation such as 也 (also) and 還 (still). In the discourse marker dictionary, the word 也 (also) is defined as a discourse marker of the *Expansion* relation. However, this word is frequent in the instances of all the four relations. In different relations, the distributions of the sentiment transitions of this word differ. In other words, the word 也 (also), which is a common word in Chinese, is not only used as a discourse marker for emphasizing the *Expansion* relation, but also has various senses in other usages.

For instance, the word 也 in (S3) is a discourse marker to denote an *Expansion* relation, but it is a particle in (S4). In fact, (S4) is an instance of the implicit *Contingency* relation. We ignore all of instances of the word 也 (also) in the following analysis since it is an outlier.

(S3) 這既是對我們工作的肯定 (This is an affirmation of our work)，也是對我們的一種鼓勵和鞭策(and also our encouragement and motivation)。

(S4) 不能放開心前行 (The mind cannot be open to forward progress)，天地也變得狹小 (the world becomes narrow)。

The word 還 (still) is another ambiguous discourse marker. Besides the *Expansion* relation defined in the dictionary, it is sometimes used to denote the *Temporal* relation, especially in the negation context, e.g., 還沒 (not yet).

The two frequent discourse markers of the *Contingency* relation, 由於 (due to) and 因為 (because) share the similar sense, and their distributions of sentiment polarity transitions are more consistent than the other markers of the *Contingency* relation.

The most frequent discourse marker of the *Comparison* class is 但 (but). The other two discourse markers 卻 (but) and 但是 (but) share the similar sense, however, their polarity distributions differ significantly. Compared to the more general marker 但 (but), the second frequent marker 卻 (but) is bolder and more critical. (S5) is an example of the marker 卻 (but). As shown in our data, the marker 卻 (but) is likely to highlight the negative sentences.

PDTB Class	Discourse Markers	#	Distribution of each type of sentiment polarity transition (%)								
			Neu Neu	Pos Neu	Neg Neu	Neu Pos	Pos Pos	Neg Pos	Neu Neg	Pos Neg	Neg Neg
Temporal	之後 (and then) in Arg1	69	‡50.72	1.45	2.90	15.94	5.80	2.90	8.70	4.35	7.25
	也 (also) in Arg2	50	†44.00	2.00	2.00	18.00	6.00	0.00	20.00	0.00	8.00
	又 (again) in Arg2	49	‡71.43	0.00	0.00	12.24	2.04	0.00	10.20	4.08	0.00
	還 (still) in Arg2	46	‡58.70	0.00	0.00	10.87	8.70	0.00	17.39	0.00	4.35
	再 (again) in Arg2	38	‡78.95	2.63	0.00	10.53	0.00	0.00	2.63	0.00	5.26
Contingency	如果 (if) in Arg1	190	†42.63	4.21	11.58	14.21	3.68	3.16	10.53	1.05	8.95
	由於 (due to) in Arg1	82	†31.71	2.44	2.44	4.88	18.29	3.66	13.41	1.22	21.95
	也 (also) in Arg2	77	20.78	0.00	1.30	20.78	19.48	0.00	11.69	2.60	†23.38
	因為 (because) in Arg1	70	†28.57	4.29	7.14	7.14	10.00	2.86	18.57	4.29	17.14
	為了 (in order to) in Arg1	62	‡50.00	14.52	1.61	6.45	9.68	1.61	8.06	6.45	1.61
Comparison	但 (but) in Arg2	176	21.59	4.55	2.84	4.55	3.41	16.48	15.91	†28.98	1.70
	卻 (but) in Arg2	85	11.76	0.00	2.35	4.71	1.18	10.59	22.35	†42.35	4.71
	而 (however) in Arg2	77	†46.75	5.19	0.00	5.19	1.30	3.90	10.39	22.08	5.19
	也 (also) in Arg2	44	†31.82	0.00	2.27	6.82	15.91	13.64	18.18	2.27	9.09
	但是 (but) in Arg2	44	15.91	4.55	0.00	0.00	2.27	25.00	11.36	†40.91	0.00
Expansion	也 (also) in Arg2	603	†43.62	1.66	1.49	15.26	19.07	1.00	7.79	0.33	9.78
	還 (still) in Arg2	231	‡50.65	2.60	0.87	11.26	14.72	0.87	9.96	0.43	8.66
	說 (say) in Arg1	206	†48.54	2.43	0.49	18.45	9.22	0.00	16.50	0.49	3.88
	並 (and) in Arg2	191	‡54.45	3.14	0.52	10.47	25.65	0.00	4.19	0.00	1.57
	也 (also) in Arg1	159	†37.11	7.55	3.14	11.95	25.16	0.63	3.77	0.63	10.06

Table 7. Five most frequent discourse makers of each PDTB class in our corpus.

(S5) 這樣觸目驚心的新型犯罪 (The new type of crime is so startling)，卻在偵破前一直沒被披露(but had never been disclosed before solved)。

The other discourser marker 但是 (but) is an emphasized version of the marker 但 (but) so that it is more likely used in the stronger polarity transitions such as *Positive*-*Negative* and *Negative*-*Positive*. In addition, the sense of the marker 而 (however) is also similar to the sense of 但 (but), but it is more frequent to be used in the neutral situations. These linguistic phenomena show that the synonyms may have different sentiment usages in the real world.

4.3 Association between discourse relation and sentiment polarity

To analyze the data at a higher level, we reorganize the sentiment transitions into several transition categories from four aspects. The details are shown in Table 8. The first aspect is *Polarity Tendency*, which classifies the transitions into three categories, including *Positive-Tendency*, *Neutral*, and *Negative-Tendency*. This aspect reflects the overall polarity of both arguments. The *Negative-Positive* transition is considered as *Positive-Tendency* because the emphasis of a Chinese sentence is usually placed in the last clause. Similarly, the *Positive-Negative* transition is considered as *Negative-Tendency*. The second aspect is *Polarity Change*, which indicates if the polarities of both arguments are opposite. Only *Negative-Positive* and *Positive-Negative* are regarded as *Opposite*. All the rest transitions are treated as *NonOpposite*. The third aspect is *Direction*, which captures the movement from the first clause to the second one. *To-Positive* stands for the transitions in which the polarity of the second clause is more positive than that of the first clause. On the other hand, *To-Negative* stands for the transitions in which the polarity of the second clause is less positive than that of the first clause. *Equal* stands for the cases in which the polarities of both clauses are identical. The last aspect is *Negativity*, which regards the polarity of an argument as binary values, i.e., *Negative* and *NonNegative*. In this way, we re-classify the nine-way sentiment polarity transitions into four transitions. In other words, both the polarity states *Neutral* and *Positive* are merged into one state *NonNegative* in this aspect. Such a binary scheme is also used in some related work, in which the negative polarity is distinguished and the rest are considered Positive (Kim and Hovy, 2004; Devitt and Ahmad, 2007). For each type of each aspect, five discourse markers that occur more than 10 times in the dataset and have the highest ratio of the corresponding type are listed in the fifth column of Table 8 as significant discourse markers.

We analyze the annotations according to the four aspects, and the results are shown in Table 9. The chi-squared test is used to test the dependency between the PDTB classes of discourse markers and each aspect of sentiment transitions. The results show that no matter whether the sentiment polarity transitions are categorized into *Polarity Tendency*, *Polarity Change*, *Direction*, or *Negativity*, the classes of discourse relations are significantly dependent on the sentiment polarities of the arguments at p=0.001.

In the aspect of *Polarity Tendency*, the ratios of *Neutral* in the *Temporal* and *Expansion* relations are 57.01% and 48.33%, respectively, which are definitely higher than those of *Contingency* and *Comparison* relations. In other words, the two arguments of *Contingency* and *Comparison* relations are less likely to be neutral. The ratio of *Negative-Tendency* of the *Comparison* relation is 46.72%. It confirms the *Comparison* relation is likely to be involved in negative statements. As shown in Table 8, three of the five significant discourse markers of *Negative-Tendency* are the synonyms of 卻 (but), which are discourse markers of the *Comparison* relation. The other two markers, 否則 (otherwise) and 因 (because), are discourse markers of the *Contingency* relation. Like the word *otherwise* in English, 否則 (otherwise) is used for introducing what bad scenario will happen if something is not done. The marker 因 (because) is not only a significant discourse marker of the category *Negative-Tendency*, but also a significant marker of *Negative-Negative* from the aspect of *Negativity*. From the real data, we find this marker is often used in bad cause-and-effect statements. (S6) is an example. The usage of the other discourse marker 因為 (because), which is a synonyms of 因 (because), is more general.

(S6) 因毛巾日久不見陽光 (Because the towel is without sunlight for a long time)，容易滋生細菌和真菌 (it is easy to breed bacteria and fungi)。

The ratio of *Opposite* of *Comparison* relation from the aspect of *Polarity Change* is 38.43%. Although it is not as high as expected, it is the highest among the four PDTB classes and much higher than those of three other classes. Compared to the other classes, *Comparison* is most likely to have a pair of opposite arguments.

Four of the five significant discourse markers of Opposite in Table 8 are the synonyms of 但 (but). Expansion relation has the highest ratio of *NonOpposite*. This matches our expectation that the *Expansion* relation is used to concatenate several events which have similar properties

Aspect	Transition Category	Sentiment polarity transitions	Explanation	Significant Discourse Markers
Polarity Tendency	Positive-Tendency	Pos-Neu, Neu-Pos, Pos-Pos, Neg-Pos	The two arguments present an overall positive polarity.	不僅...也 (not only... also), 終於 (finally) , 既...又 (now that...), 只要 ... 就 (as long as...), 近年 (recently)
	Neutral	Neu-Neu	Both arguments are neutral.	然後 (and then), 因此 (hence) , 最後 (at the end), 故 (so) , 以及 (as well as)
	Negative-Tendency	Pos-Neg, Neg-Neu, Neu-Neg, Neg-Neg	The two arguments present an overall negative polarity.	否則 (otherwise), 卻 (but), 可是 (but), 但是 (but), 因 (because)
Polarity Change	Opposite	Neg-Pos, Pos-Neg	The polarities of both arguments are opposite.	但是 (but), 雖然...但 (although...) , 但 (but), 卻 (but), 不過 (but)
	NonOpposite	Neu-Neu, Pos-Neu, Neg-Neu, Neu-Pos, Pos-Pos, Neu-Neg, Neg-Neg	The polarities of both arguments are not opposite.	或 (or), 像 (as), 而且 (moreover), 如果...會 (if ... may), 表示 (say)
Direction	To-Positive	Neg-Neu, Neg-Pos, Neu-Pos	The second argument is less negative than the first one.	終於 (finally), 雖然 ... 但 (although...) , 近年 (recently), 只要 ... 就 (as long as...) , 看來 (seem...)
	Equal	Neg-Neg, Neu-Neu, Pos-Pos	Both arguments are the same polarity value.	不僅...更 (Not only... even), 最後 (at the end), 並且 (in addition), 故 (so), 既...也 (now that...)
	To-Negative	Pos-Neu, Pos-Neg, Neu-Neg	The second argument is less positive than the first one.	卻 (but), 但是 (but), 可是 (but), 否則 (otherwise), 即使...也 (even if...)
Negativity	NonNegative-NonNegative	Neu-Neu, Neu-Pos, Pos-Neu, Pos-Pos	Both arguments are not negative.	以及 (as well as), 未來 (in the future), 以便 (in order to), 並且 (in addition), 然後 (and then)
	NonNegative-Negative	Neu-Neg, Pos-Neg	The first argument is not negative while the second argument is negative.	卻 (but), 否則 (otherwise), 但是 (but), 即使...也 (even if...), 可是 (but)
	Negative-NonNegative	Neg-Neu, Neg-Pos	The first argument is negative while the second argument is not negative.	雖然...但 (although...), 但是 (but), 不過 (but), 終於 (finally), 但 (but)
	Negative-Negative	Neg-Neg	Both arguments are negative.	甚至 (even), 卻 (but), 因 (because), 如果...將 (if... may), 但是 (but)

Table 8: Aspects of sentiment transition.

PDTB Class	#	Polarity Tendency (%)			Polarity Change (%)		Direction(%)			Negativity (%)			
		Pos Tend	Neutral	Neg Tend	Oppo	Non Oppo	To Pos	Eq.	To Neg	NonNeg-NonNeg	NonNeg-Neg	Neg-NonNeg	Neg-Neg
Tem	849	23.79	57.01	19.20	3.42	96.58	20.85	63.84	15.31	78.45	13.78	4.48	3.30
Con	1,598	30.16	35.42	34.42	4.13	95.87	21.90	60.95	17.15	63.27	13.45	8.20	15.08
Com	929	30.14	23.14	46.72	38.43	61.57	26.80	30.89	42.30	37.57	39.61	18.19	4.63
Exp	4,262	33.88	48.33	17.79	1.22	98.78	16.75	71.89	11.36	81.63	8.49	2.51	7.37

Table 9: Statistics of sentiment transition for each PDTB class over the corpus annotated by human.

from certain perspective.

The ratio of *To-Negative* of *Comparison* relation from the aspect of *Direction* in Table 9 is 42.30%, which is significantly higher than the ratios of *To-Negative* of the other classes. This also confirms the *Comparison* relation is likely to be used to express critical opinions. Furthermore, the ratio of *Equal* of *Comparison* relations is much lower than those of other classes. This result shows the *Comparison* relation is more involved in sentiment polarity transitions.

The *Negativity* aspect in Table 9 also shows the *NonNegative-Negative* is more likely to happen than the *Negative-NonNegative* in all relations. This statistics reflects a particular phenomenon "good words ahead" in Chinese. That is, speakers tend to express a negative opinion after kind words.

The sentiment polarity flips in the instances of the two categories *Negative-NonNegative* and *NonNegative-Negative*. However, the significant discourse markers of the two categories are very different. In spite of the general marker 但是 (but), the discourse markers 卻 (but), 否則 (otherwise), 即使...也 (even if...), and 可是 (but) are often used in *NonNegative-Negative*, which usually results a negative remark. On the other hand, the discourse markers 雖然...但 (although...), 不過 (but), 終於 (finally), and 但 (but) are often used in *Negative-NonNegative*, which usually results a positive remark. For example, the discourse marker 終於 (finally), which is a discourse marker of the *Temporal* relation, is usually used when an event successfully accomplished after twists and turns such as (S7).

(S7) 歷經多次磨難的國產手機巨頭波導 (Domestic mobile phone giant Ningbo Bird after many tribulations)，終於成功轉戰汽車行業 (finally successfully fought in the automotive industry)。

5 Conclusion

To investigate the discourse relation and the sentiment polarity of Chinese discourse markers, we construct a moderate-sized corpus based on the Chinese part of ClueWeb09. In this paper, our annotation scheme and the analysis of the annotation results are shown. Total 7,638 instances are annotated by native speakers. The discourse relation distribution of the annotated data is comparable to the distribution of the well-known English discourse corpus PDTB 2.0. Through the data analysis, we validate certain human intuitions in Chinese language. Near half of instances are in neutral sentiment while the *Comparison* relation is most likely to be involved in negative sentiment. Furthermore, the high sentiment dependency between the last clause and the whole sentence is validated in the data.

The data shows the significant association between the discourse relation and the sentiment polarity. The arguments of a *Comparison* relation or a *Contingency* relation are more likely to be involved in expressing sentiment. Moreover, the *Comparison* relation often occurs in the sentences with sentiment polarity transitions, and frequently occurs in the instances with the negative sentiment. On the other hand, the arguments of the *Temporal* and the *Expansion* relations are relatively objective. The behavior of word choice between synonyms is also observed in the data. Each synonym of a sense may have its own usage in expressing sentiment.

This paper points out the ambiguities of the discourse markers in Chinese. That is, a marker may suggest more than one discourse relation. Besides, words may have both the functions of discourse connectives and non-discourse ones in their surface forms. These two issues make the interpretation of Chinese discourse markers more challenging. Determination of their correct uses and disambiguation of their discourse functions will be investigated in the future.

Acknowledgments

This research was partially supported by Excellent Research Projects of National Taiwan University under contract 102R890858 and 2012 Google Research Award.

References

Lynn Carlson and Daniel Marcu. 2001. Discourse Tagging Reference Manual. http://www.isi.edu/~marcu/discourse/tagging-ref-manual.pdf

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. *RST Discourse Treebank*. Linguistic Data Consortium, Philadelphia.

Shou-Yi Cheng. 2006. *Corpus-Based Coherence Relation Tagging in Chinese Discourse*. Master's Thesis, National Chiao Tung University, Hsinchu, Taiwan.

Xianghui Cheng and Xiaolin Tian. 1989. *Xian dai Han yu* (*現代漢語*), San lian shu dian (三聯書店), Hong Kong.

Ann Devitt and Khurshid Ahmad. 2007. Sentiment polarity identification in financial news: a cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (*ACL 2007*), pages 984-991, Prague, Czech Republic.

Hugo Hernault, Helmut Prendinger, David A. duVerle, and Mitsuru Ishizuka. 2010. HILDA: A Discourse Parser Using Support Vector Machine Classification. *Dialogue and Discourse*, 1(3): 1-33.

Hen-Hsen Huang and Hsin-Hsi Chen. 2011. Chinese discourse relation recognition. In *Proceedings of the 5th International Joint Conference on Natural Language Processing* (*IJCNLP 2011*). pages 1442-1446, Chiang Mai, Thailand.

Hen-Hsen Huang and Hsin-Hsi Chen. 2012a. Contingency and comparison relation labeling and structure prediction in Chinese sentences. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (*SIGDIAL 2012*), pages 261-269, Seoul, South Korea.

Hen-Hsen Huang and Hsin-Hsi Chen. 2012b. An Annotation System for Development of Chinese Discourse Corpus. In *Proceedings of the 24th International Conference on Computational Linguistics* (*COLING 2012*)*: Demonstration Papers*, pages 223-230, Mumbai, India.

Ben Hutchinson. 2004. Acquiring the meaning of discourse markers. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), pages 684-691, Barcelona, Spain.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics* (*COLING-04*), pages 1367-1373, Geneva, Switzerland.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2009*), pages 343-351.

Shuxiang Lu. 2007. *Eight Hundred Words of The Contemporary Chinese (Xian dai Han yu Ba bai Ci)*, China Social Sciences Press.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2): 1-135.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2007. *The Penn Discourse Treebank 2.0 Annotation Manual*. The PDTB Research Group.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the 6th Language Resources and Evaluation Conference* (*LREC 2008*), pages 2961-2968, Marrakech, Morocco.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (*HLT/NAACL 2003*), pages 149-156, Edmonton, Canada.

Florian Wolf and Edward Gibson. 2005. Representing Discourse Coherence: A Corpus-Based Analysis. *Computational Linguistics*, 31(2): 249-287.

Chi-Hsin Yu, Yi-jie Tang and Hsin-Hsi Chen. 2012. Development of a web-scale Chinese word N-gram corpus with parts of speech information. In *Proceedings the 8th International Conference on Language Resources and Evaluation* (*LREC 2012*), pages 320-324, Istanbul, Turkey.

Zhi-Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew Lim Tan. 2010. Predicting discourse connectives for implicit discourse relation recognition. In *Proceedings of the 23rd International Conference on Computational Linguistics* (*COLING 2010*)*: Posters*, pages 1507-1514.

Yuping Zhou and Nianwen Xue. 2012. PDTB-style discourse annotation of Chinese text. In *Proceedings the 50th Annual Meeting of the Association for Computational Linguistics* (*ACL 2012*), pages 69-77, Jeju, South Korea.