Web-Based Analysis of Chinese Discourse Markers for Opinion Mining

Hen-Hsen Huang, Chi-Hsin Yu, Tai-Wei Chang, Cong-Kai Lin, and Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, 10617 Taiwan
{hhhuang, jsyu, twchang, cklin}@nlg.csie.ntu.edu.tw, hhchen@ntu.edu.tw

***Abstract*—Discourse markers not only express some sorts of relations between two arguments, but also entail sentiment information. In this paper, we investigate the associations between the relation type and the sentiment polarity of Chinese discourse markers based on a web scale corpus. We present an approach to mining information from a large scale corpus, show the polarity distributions of sentences under various relation types, and interpret the data from various aspects. Based on the massive amount of data from the Internet, certain language phenomena are shown.**

Keywords—Chinese Discourse Analysis; Discourse Relation Labeling; Sentiment Analysis

I. INTRODUCTION

Discourse markers, also called connectives in Penn Discourse Treebank [1], play a crucial role in recognizing the discourse relation between arguments. A connective joins two discourse units such as phrases, clauses, or sentences together. For example, the word "because" is a connective that indicates a Contingency relation between two clauses.

In the work of the Penn Discourse Treebank 2.0 annotation [1], experts 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 also defined three levels of sense hierarchy for the connectives. The four classes on the top level are Temporal, Contingency, Comparison, and Expansion. The sense of a connective denotes how its two arguments cohere. In other words, a connective presents the relation of its two arguments.

In addition to the discourse relation, Hutchinson pointed out the properties of a discourse marker from different aspects such as veridicality and sentiment polarity [2]. Veridicality examines whether both the two arguments are true or not, and the sentiment polarity denotes the sentiment transition of the two arguments of a discourse marker.

In Chinese domain, our previous work addressed the interaction between the sentiment polarity and the discourse structure in Chinese [3-4]. (S1) shows an example that the sentiment in the first clause is positive, and the sentiment in the second clause is negative. The discourse relation between these two clauses is Comparison.

(S1) 回答這個問題很容易，解決卻很難。 ("It is easy to answer this problem, but it is difficult to solve it.")

As the PDTB 2.0 annotation manual suggests [5], a Comparison relation is happened to contrast the differences between the two arguments. Therefore, it is expected that the two arguments of a Comparison relation are 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 likely to belong to the same polarity states (e.g., Positive-Positive or Neutral-Neutral).

Discourse relation recognition [6-8] and sentiment analysis [9] are two hot topics that have rapid growth in these years. In this paper, we analyze the association between discourse relations and sentiment polarities based on a large scale corpus. To obtain the information on these two aspects for a given sentence, a dictionary of Chinese discourse markers is consulted and a web corpus approach are adopted. In the Chinese discourse marker dictionary, the sense of each marker is listed. We determine the sentiment polarity of each discourse marker using a web scale corpus, i.e., the ClueWeb. To verify the reliability of the results mined from the web scale corpus, we compare the Chinese portion of the ClueWeb09 with a Chinese balanced corpus, i.e., the Academic Sinica Balanced Corpus (ASBC) [10]. To the best of our knowledge, this is the first attempt that addresses the correlation between Chinese discourse relation and sentiment analysis using the large-scale web corpus.

The rest of this paper is organized as follows. In Section II, we introduce the dictionary of Chinese discourse markers, the ClueWeb dataset and the ASBC corpus. In Section III, the methodology to measure the sentiment score for each discourse marker is introduced and evaluated. In Section IV, we show some results mined from the ClueWeb and provide the in-depth analyses on the discourse relation and sentiment polarity. Finally, we conclude the remarks.

II. LINGUISTIC RESOURCES

The huge amount of data in the web scale corpus entails the human knowledge on the use of languages. To explore such information, we adopt a public available Chinese Web POS tagged corpus [15] to extract the needed information. This

978-1-4799-4143-8/14 $31.00 © 2014 IEEE
DOI 10.1109/WI-IAT.2014.84

Chinese POS-tagged corpus was developed based on the ClueWeb09 dataset, where Chinese material is the second largest. It contains 9,598,430,559 POS-tagged sentences in 172,298,866 documents. A small corpus sampled from the ClueWeb09 with human annotation [4] and the Academic Sinica Balanced Corpus [10] are used as our reference corpora. In the ASBC, there are about 5 million words from traditional Chinese articles. The ClueWeb is much larger, but contains more noise, while the ASBC is much cleaner, but is smaller.

Discourse markers are the fundamental elements of this work. We adopt a Chinese discourse marker dictionary which is developed and extended by experts [11-13]. Table I shows the overview of the discourse marker dictionary. The dictionary consisted of 808 Chinese discourse markers (single words or paired words) categorized as ten types [11] [14], shown in the first column. Type is the first dimension of Chinese discourse markers. Similar to the sense that is annotated in the PDTB, it is the type of a discourse relation. Although the languages are different, the ten types defined in the Chinese discourse marker dictionary can be mapped into the four classes annotated in the PDTB as shown in the second column. For example, the four types, including Cause-effect, Condition, Generalization, and Purpose, belong to the class Contingency in the PDTB, and the type Temporal Sequence is identical to the class Temporal in the PDTB. In the rest of this paper, we denote the ten type scheme as CT10.

In addition to the types of discourse relations, we classify the markers into three groups of scopes shown in the third 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. The fourth column lists the number of discourse markers of each scope, and the fifth column gives some samples.

III. Analysis Methodology

We select the two-clause sentences that contain exact one discourse marker from the Chinese portion of the ClueWeb09 dataset. As a result, a total of 43,000,471 sentences meeting the constraints are collected. Since no discourse and sentiment tags are available in the raw ClueWeb and in the ASBC, we apply a sentiment tagger to predict the sentiment polarities for the extracted sentences and label the discourse relation to each instance by looking up the discourse marker dictionary.

TABLE I. Overview of a Chinese discourse marker dictionary

CT10 Type	PDTB Class	Scope	# Markers	Examples
Similarity	Expansion	Single word	22	另外 (besides)
		Intra-sentential	19	一方面…一方面 (on the one hand ... on the other hand)
		Inter-sentential	5	首先…再者 (first ... second)
Temporal Sequence	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)
Alternative	Expansion	Single word	10	抑或 (or)
		Intra-sentential	32	不是…而是 (not ... but)
		Inter-sentential	9	或…或許 (or ... perhaps)
Elaboration	Expansion	Single word	25	不只 (not only)
		Intra-sentential	55	不只…也 (not only ... also)
		Inter-sentential	12	不只…不只 (not only ... not only)
Violated Expectation	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)
Cause-effect	Contingency	Single word	13	因爲 (because)
		Intra-sentential	36	因…而 (because ... then)
		Inter-sentential	11	既然…於是 (since ... then)
Condition	Contingency	Single word	21	如 (if)
		Intra-sentential	93	如…則 (if then)
		Inter-sentential	3	至少…不然 (at least ... otherwise)
Generalization	Contingency	Single word	28	假設 (suppose)
		Intra-sentential	51	凡…可 (any ... can)
Example	Expansion	Single word	120	例如 (such as)
Purpose	Contingency	Single word	5	以免 (in order to avoid)

A. A Lexicon-based Method for Sentiment Analysis

In this stage, we have to assign a sentiment score to each clause. Various sentiment analyses from lexicon-based approaches [16] to syntax/dependency-based approaches [17-18] have been proposed. The lexicon-based approach, which determines sentiment polarity based on lexicons, is simple, but the lexicon coverage, context-sensitive phenomena, and biases of the uses of positive and negative words [19] may be problems. The syntax-based and the dependency-based approaches, which consider the structural relations or dependency relations of linguistic elements to determine sentiment polarity, may capture more semantics, but error propagation from parsing may be a problem.

In this work, an efficient lexicon-based method is adopted for sentiment analysis. A public available sentiment dictionary, NTUSD [20], is referred to determine the polarity of a word. The NTUSD contains 43,805 Chinese sentiment words, of which 21,056 words are positive and 22,749 words are negative. Those words outside of the NTUSD are regarded as neutral. Given a clause, our algorithm matches all the sentiment words and summarizes their scores. The summation is further normalized with respect to the length of the clause. The scores of a positive and a negative sentiment words are set to 1 and -1, respectively. The score of a sentiment word will be negated if a negation is found before the word. The longest matching is taken when multiple matches occur.

To avoid the phenomena of positivity bias in writing [21-23], we average the sentiment scores of the first and the second arguments of all the sentences as new neutral scores, and offset all the scores to the new base. For the analysis, we group the sentiment scores by three levels of categories, i.e., individual discourse markers, the CT10 types in the discourse marker dictionary, and the PDTB four classes. The scores belonging to the same category are averaged in proportion to their occurrences.

B. Evaluation

In order to validate of the results automatically mined from the ClueWeb, we compare the mined data to the human annotate corpus as ground truth [4]. The human annotated corpus contains a total of 7,638 instances that are also randomly selected from the ClueWeb09 and annotated by 87 native speakers. We evaluate our lexicon-based sentiment tagger with the annotated instances. The tagging performance is reported in Table II. The accuracy of classifying Negative clauses from Non-Negative ones is the highest among all classification tasks. Therefore, we further analyze the data in the scheme of Negative/Non-Negative in Section IV.

TABLE II. PERFORMANCE OF THE SENTIMENT TAGGER

Classification	Accuracy
Positive vs. NonPositive	74.65%
Neutral vs. NonNeutral	61.67%
Negative vs. NonNegative	81.65%
Average	72.66%

C. Comparison of the Three Corpora

Besides the performance evaluation of the sentiment tagger, we also verify whether the instances from the ClueWeb can reflect the language uses in real world. We compare the distribution of the mined instances with the distribution of the balanced corpus, i.e., ASBC. In Table III, the distributions of the CT10 types of discourse markers from the ASBC, the ClueWeb, and the human annotated corpus are shown. The symbol # denotes the occurrences of instances, % denotes the percentage of the occurrences, and R denotes the ranking order of occurrences. The three sets significantly differ in size, but share a similar type distribution. Similarity Contrast is the most frequent type of discourse markers in both corpora, and Example is the second frequent type. Both these two types belong to the Expansion relation in the PDTB. To verify the similarity of the distributions of these three sets, we calculate the spearman's ρ values between the pairwise rankings. The ρ value of the rankings between the ASBC and the ClueWeb is 0.91, the ρ value of the rankings between the ClueWeb and the Human is 0.96, and the ρ value of the rankings between the ASBC and the Human is 0.87. The high ρ values indicate the distributions in the three sets are similar at the 0.01 level of significance. The distributions of the discourse markers in the four PDTB classes are shown in Table IV. The three distributions are also similar and have the same rankings. Although the data in the ClueWeb is messier, this experiment shows the discourse marker distributions of the ASBC, the

TABLE III. CT10 TYPE DISTRIBUTIONS OF THE ASBC, THE CLUEWEB, AND THE HUMAN ANNOTATED CORPUS

CT10 Type	ASBC			ClueWeb			Human		
	#	%	R	#	%	R	#	%	R
Similarity Contrast	1516	18.83	1	35,432,592	17.52	1	1,566	19.87	1
Temporal Sequence	957	11.88	4	20,074,351	9.93	6	840	10.66	6
Alternative	626	7.77	7	24,239,161	11.98	5	872	11.06	5
Elaboration	795	9.87	5	26,106,008	12.91	4	1,086	13.78	3
Violated Expectation	1,239	15.39	3	26,364,871	13.04	3	1,220	15.48	2
Cause-effect	790	9.81	6	12,784,488	6.32	8	532	6.75	8
Condition	465	5.77	8	17,828,731	8.81	7	560	7.11	7
Generalization	157	1.95	9	4,603,592	2.28	9	165	2.09	9
Example	1,492	18.53	2	31,546,185	15.60	2	924	11.72	4
Purpose	16	0.20	10	3,280,123	1.62	10	116	1.47	10

TABLE IV. PDTB CLASS DISTRIBUTION OF THE ASBC, THE CLUEWEB, AND THE HUMAN ANNOTATED CORPUS

PDTB Class	ASBC			ClueWeb			Human		
	#	%	*R*	#	%	*R*	#	%	*R*
Temporal	957	11.88	4	20,074,351	9.93	4	840	10.66	4
Contingency	1,428	17.73	2	38,496,934	19.03	2	1,373	17.42	2
Comparison	1,239	15.39	3	26,364,871	13.04	3	1,220	15.48	3
Expansion	4,428	55.00	1	117,323,946	58.01	1	4,448	56.44	1

ClueWeb, and the human- annotated corpus are similar from the aspects of CT10 type and PDTB class. The result also shows that the sentences sampled from the ClueWeb are representative.

IV. RESULTS AND DISCUSSION

The mined results are analyzed on the levels of specific discourse markers and the classes of discourse relations. The chi-squared test is applied to validate the correlation between the discourse relation and the sentiment polarity.

A. Frequent Discourse Markers

The top discourse markers in the ClueWeb are shown in Table V. The five most frequent discourse markers for each of the four PDTB classes are listed in the order of Temporal, Contingency, Comparison, and Expansion. In each row of the table, the discourse marker and the distribution of its nine sentiment polarity transitions are given. Recall that there are three polarities, i.e., positive, neutral, and negative. The major sentiment polarity transition of each discourser maker is labelled with the symbol †.

The most frequent discourse marker in all classes is the Single word 也 (also) that belongs to the PDTB class Expansion and the CT10 type Similarity Contrast. The major polarity transition of 也 (also) is Positive-Positive. It accounts the cases that the polarities of both arguments of 也 (also) are positive. (S2) is an example of this discourse marker with the Positive-Positive transition. (S3) forms a Negative-Negative transition. The occurrences of the other most frequent transitions, Negative-Negative (14.72%) and Neutral-Neutral (14.63%), are very close to the occurrence of Positive- Positive (14.88%). The common property of these three transitions is that sentiment polarities of both clauses are identical, i.e., no polarity changes.

(S2) 當對方被這些東西逗得很開心時 (When the counterpart is amused very happy with these things)，她也會開始期待下一次的驚喜 (she will start to look forward to the next surprise)。

(S3) 不利健康 (Not conductive to health)，也不利和諧社會的構建啊 (and not conducive to the establishment of a harmonious society, too)。

The two most frequent discourse markers of the class Contingency are 如果 (if) and 爲了 (in order to). They are likely to occur with the sentiment polarity transition, Positive-Positive. The discourse marker 爲了 (in order to) is an interesting case because it is usually used in the situation of persuasion [24]. The typical pattern is “爲了 Arg1，請 Arg2” (In order to Arg1, please Arg2). Therefore, both the two arguments are likely to be positive. (S4) is an example of this word used for persuasion.

(S4) 爲了方便我們及時通知您獲獎資訊 (In order to facilitate our immediate notification of your winning information)，請務必填寫正確的資訊 (please be sure to fill in the correct information)。

The most frequent discourse marker of the class Comparison is 而 (however). We expect the two arguments of a Comparison relation are polar opposites. Unexpectedly, the major sentiment polarity transition of 而 (however) is Negative-Negative, rather than the transitions with opposite polarity. This word shows an interesting case in Chinese. It is listed in the discourse marker dictionary as the Comparison relation. However, it is sometimes used as the marker of Expansion for the sense of “moreover”. For instance, the word 而 (however) in (S5) plays the Comparison relation. In contrast, 而 (however) in (S6) is a discourse marker denoting an Expansion relation of the two clauses. This word reveals the ambiguity of discourse markers. It has to be resolved.

(S5) 請視它們爲經典 (Please treat them as classic)，而不是陳舊 (but not trite)。

(S6) 很多企業都在尋求「變綠」的機會 (A lot of companies are looking for the opportunity of “greenizing”)，而你正好可以爲他們提供這樣的機會(and you can just provide them with this opportunity)。

The rest four top discourse markers of the class Comparison are synonyms that have the sense of “but”. However, the polarity distribution of the marker 卻 (but) is different from those of the other markers. (S7) is an example of the marker 卻 (but). Compared to the more general maker 但 (but), the marker 卻 (but) is more critical. As shown in our data, the marker 卻 (but) is likely to highlight the negative situations. These linguistic phenomena show that the synonyms may have different usages in the real world.

TABLE V. FIVE MOST FREQUENT DISCOURSE MAKERS OF EACH PDTB CLASS IN THE CLUEWEB.

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*
現在 now	†15.60	10.37	7.31	11.94	12.87	9.56	11.76	10.81	9.79
然後 then	†19.19	11.03	11.73	10.41	11.94	10.74	9.19	6.71	9.06
終於 finally	14.37	8.26	5.74	†17.38	15.02	9.67	12.46	9.69	7.41
曾 ever	†22.53	7.37	6.36	11.96	7.00	6.11	18.17	10.38	10.13
目前 now	7.40	6.84	4.54	8.46	16.02	12.76	10.82	†20.69	12.47
如果 if	3.94	5.44	6.34	6.71	†36.75	18.26	5.06	9.50	8.00
爲了 for	3.72	7.71	2.53	4.49	†34.17	5.50	6.02	28.35	7.51
可能 perhaps	7.13	6.97	6.12	9.43	14.61	12.06	11.11	16.02	†16.56
因爲 because	7.51	4.33	4.75	11.06	11.01	9.66	†20.47	14.72	16.49
所以 so	9.27	3.34	2.53	18.27	9.16	6.25	†33.99	9.84	7.35
而 however	12.74	5.21	8.55	6.96	9.42	11.92	10.01	13.91	†21.27
但 but	5.07	4.60	5.01	5.67	12.59	14.65	9.21	†23.06	20.12
卻 but	11.10	5.50	6.39	8.43	8.04	10.03	15.42	16.70	†18.36
但是 but	4.96	6.48	6.63	6.13	13.02	13.81	8.54	†21.02	19.40
不過 but	11.96	9.03	7.11	8.57	11.06	10.46	11.94	†15.13	14.74
也 also	14.63	5.77	5.25	10.14	†14.88	9.79	12.30	12.53	14.72
還 still	†18.55	8.22	8.44	12.17	9.92	8.04	14.17	10.08	10.42
更 moreover	8.47	5.01	3.57	19.34	†26.97	13.05	6.21	8.79	8.59
或 or	10.72	5.79	3.11	11.71	14.65	14.22	8.21	†18.83	12.77
並 and	5.47	5.10	4.83	7.97	24.54	†25.23	4.80	10.01	12.05

(S7) 這樣觸目驚心的新型犯罪 (The new type of crime so startling)，卻在偵破前一直沒被披露 (but had never been disclosed before solved) 。

From the real data, we also find some examples have pragmatically opposite arguments. This problem is challenging for the lexical-based sentiment tagger. For instance, both arguments in (S8) are semantically positive. However, the adjective 年輕 (young), which is defined as positive in NTUSD, may imply the sense of lacking in experience or unskilled when it is applied to modify a sportsman.

(S8) 他很年輕 (He is young)，但已經是世界上最棒的足球運動員之一(but he has been one of best soccer players in the world) 。

The major sentiment polarity transitions of Temporal discourse markers are Neutral-Neutral. The reason is 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 sentiment words are relatively rare. One exception is the discourse marker 終於 (finally) that is likely to form the transitions of Neutral-Positive and Positive-Positive. From the real data, we find that the maker 终於 (finally) is usually used when an event successfully accomplished in the end. (S9) is an example.

(S9) 聚沙成塔 (Many little drops make an ocean)，终於化腐朽爲神奇 (and finally the corruptible is transformed into be miraculous)。

B. Association between discourse relation and sentiment polarity

To analyze the data at a higher level, we reorganize the sentiment transitions into three transition categories – say, Polarity Tendency, Polarity Change, and Negativity. The details of each aspect are summarized in Table IV.

In short, the aspect of Polarity Tendency measures the overall polarity of both arguments. The aspect of Polarity Change indicates if the two arguments in a sentence are polar opposites. The last aspect, Negativity, 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. As mentioned in Section III C, the lexicon-based sentiment tagger produces most reliable prediction in this classification task.

We count the frequency of each category for all the discourse markers, and group them into the four PDTB classes. The results are shown in Table VII. The chi-squared test is used to test the dependency between the types and the classes of discourse markers, and the categories of sentiment transitions. The results show that no matter whether the sentiment polarity transitions are categorized into Polarity Tendency, Polarity Change, or Negativity, the senses of discourse markers are significantly dependent on the sentiment polarities of the arguments at p=0.000001.

TABLE VI. ASPECTS OF SENTIMENT TRANSITION

Aspect	Transition Category	Sentiment polarity transitions	Explanation
Polarity Tendency	Positive Tendency	Pos-Neu, Neu-Pos, Pos-Pos, Neg-Pos	The two arguments present an overall positive polarity.
	Neutral	Neu-Neu	Both arguments are neutral.
	Negative Tendency	Pos-Neg, Neg-Neu, Neu-Neg, Neg-Neg	The two arguments present an overall negative polarity.
Polarity Change	Opposite	Neg-Pos, Pos-Neg	The polarities of both arguments are opposite.
	Non Opposite	Neu-Neu, Pos-Neu, Neg-Neu, Neu-Pos, Pos-Pos, Neu-Neg, Neg-Neg	The polarities of both arguments are not opposite.
Negativity	NonNegative-NonNegative	Neu-Neu, Neu-Pos, Pos-Neu, Pos-Pos	Both arguments are not negative.
	NonNegative-Negative	Neu-Neg, Pos-Neg	The first argument is not negative while the second argument is negative.
	Negative-NonNegative	Neg-Neu, Neg-Pos	The first argument is negative while the second argument is not negative.
	Negative-Negative	Neg-Neg	Both arguments are negative.

TABLE VII. STATISTICS OF SENTIMENT TRANSITION FOR EACH PDTB CLASS

PDTB Class	Polarity Tendency (%)			Polarity Change (%)		Negativity (%)			
	Pos. Tend	*Neu.*	*Neg. Tend.*	*Oppo.*	*Non Oppo.*	*NonNeg-NonNeg*	*NonNeg-Neg*	*Neg-NonNeg*	*Neg-Neg*
Temporal	41.18	15.67	43.14	19.60	80.40	48.58	28.25	13.86	9.31
Contingency	47.10	8.79	44.10	26.36	73.64	42.72	25.29	19.59	12.41
Comparison	35.40	9.56	55.04	27.22	72.78	34.00	32.12	16.89	16.99
Expansion	43.63	14.85	41.03	22.04	77.47	47.45	22.85	17.75	11.46

C. Discussion

The high ratios of NonOpposite of Temporal and Expansion relations from the aspect of polarity change show that the polarity states of the two arguments of a Temporal relation and an Expansion relation tend to be the same. Most instances of Temporal relations describe the fact along the temporal sequence. Only a few markers such as 當時…現在 (at that time ... now) involve the hint to highlight the difference between situations in different moments. Expansion relation, the second to Temporal relation in Table VII, also has a high ratio of NonOpposite. This matches our expectation that the Expansion relation is used to concatenate several events which have similar properties from certain perspective.

The ratio of Opposite of Comparison relation from the aspect of polarity change is 27.22% shown in Table VII. Although it is not as high as expected, it is the highest among the four PDTB classes. Compared to the other classes, Comparison is more likely to have a pair of opposite arguments. As we discussed in Section IV A, some Comparison instances have pragmatically opposite arguments. Unfortunately, current sentiment tagger is hard to measure the polarity on the pragmatic level.

In the aspect of Polarity Tendency, the ratios of Neutral in the Temporal and Expansion relations are 15.67% and 14.85%, respectively, which are definitely higher than those of Contingency and Comparison relations. Similarly, the two arguments of Contingency and Comparison relations are less likely to be neutral. The ratio of Negative Tendency of the Comparison relation is 55.04% which confirms the Comparison relation is likely to be involved in negative statements.

The Negativity aspect in Table VII also shows the NonNegative to Negative is more likely to happen than the Negative to 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.

V. CONLUSION

This paper investigates the association between the discourse relation and the sentiment polarity of Chinese discourse markers on huge amount of data from the Internet. We show an approach to mine meaningful information from a large-scale corpus and discuss the language phenomena based on the massive amount of data. On the one hand, the arguments of the Temporal and the Expansion relations are likely to be neutral. On the other hand, more sentiment words are involved in the arguments of the Comparison and the Contingency relations. Furthermore, the two arguments of a Comparison relation are more likely to be polar opposites. The behavior of word choice between synonyms is also observed in the data. We also confirm the ambiguity of the discourse markers, i.e., a marker may suggest more than one discourse relation. Disambiguation of such markers will be investigated in the future.

ACKNOWLEDGMENT

This research was partially supported by Ministry of Science and Technology, Taiwan, under the grant 102-2221-E-002-103-MY3 and 2012 Google Research Award.

REFERENCES

[1] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber, "The Penn discourse treebank 2.0," in Proc. 6th Language Resources and Evaluation Conf. Marrakech, Morocco, 2008, pp. 2961-2968.

[2] B. Hutchinson, "Acquiring the meaning of discourse markers," in Proc. 42nd Annu. Meeting of the Association for Computational Linguistics. Barcelona, Spain, 2004, pp. 684-691.

[3] H. H. Huang and H. H. Chen, "Contingency and comparison relation labeling and structure prediction in Chinese sentences," in Proc. 13th Annu. Meeting of the Special Interest Group on Discourse and Dialogue. Seoul, South Korea, 2012, pp. 261-269.

[4] H. H. Huang, C. H. Yu, T. W. Chang, C. K. Lin, and H. H. Chen, "Analyses of the association between discourse relation and sentiment polarity with a Chinese human-annotated corpus," in Proc. ACL 2013 7th Linguistic Annotation Workshop & Interoperability with Discourse. Sofia, Bulgaria, 2013, pp. 70-78.

[5] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber, "The Penn discourse treebank 2.0 annotation manual," The PDTB Research Group, 2007.

[6] R. Soricut and D. Marcu, "Sentence level discourse parsing using syntactic and lexical information," in Proc. Human Language Technology Conf. of the North American Chapter of the Association for Computational Linguistics. Edmonton, Canada, 2003, pp. 149-156.

[7] H. Hernault, H. Prendinger, D. A. duVerle, and M. Ishizuka, "Hilda: a discourse parser using support vector machine classification," Dialogue and Discourse, vol. 1, no. 3, pp. 1-33, 2010.

[8] H. H. Huang and H. H. Chen, "Chinese discourse relation recognition," in Proc. 5th International Joint Conf. on Natural Language Processing. Chiang mai, Thailand, 2011, pp. 1442-1446.

[9] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1-135, 2008.

[10] C. R. Huang and K. J. Chen, "A Chinese corpus for linguistics research," in Proc. 14th International Conf. on Computational Linguistics. France, 1992, pp. 1214-1217.

[11] X. Cheng and X. Tian, "Xian dai han yu (現代漢語)," San lian shu dian, Hong Kong, 1989.

[12] S. Y. Cheng, "Corpus-based coherence relation tagging in Chinese discourse," Master's Thesis, National Chiao Tung University, Taiwan, 2006.

[13] S. Lu, "Eight hundred words of the contemporary Chinese (現代漢語八百詞), china social sciences pres," , S.: Eight Hundred Words of The Contemporary Chinese (現代漢語八百詞), China Social Sciences Press, 2007.

[14] F. Wolf and E. Gibson, "Representing discourse coherence: a corpus-based analysis," Computational Linguistics, vol. 31, no. 2, pp. 249-287, 2005.

[15] C. H. Yu, Y. J. Tang, and H. H. Chen, "Development of a web-scale Chinese word n-gram corpus with parts of speech information," in Proc. the 8th International Conf. on Language Resources and Evaluation. Istanbul, Turkey, 2012, pp. 320-324.

[16] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," Computational Linguistics, vol. 37, no. 2, pp. 267-307, 2011.

[17] L. W. Ku, T. H. Huang, and H. H. Chen, "Using morphological and syntactic structures for Chinese opinion analysis," in Proc. Conf. on Empirical Methods in Natural Language Processing. Singapore, 2009, pp. 1260-1269.

[18] L. W. Ku, T. H. Huang, and H. H. Chen, "Predicting opinion dependency relations for opinion analysis," in Proc. 5th International Joint Conf. on Natural Language Processing. Chiang Mai, Thailand, 2011, pp. 345-353.

[19] P. Rozin, L. Berman, and E. Royzman, "Biases in use of positive and negative words across twenty natural languages," Cognition and Emotion, vol. 24, no. 3, pp. 536-548, 2010.

[20] L. W. Ku and H. H. Chen, "Mining opinions from the web: beyond relevance retrieval," Journal of American Society for Information Science and Technology, vol. 58, no. 12, pp. 1838-1850, 2007.

[21] A. A. Augustine, M. R. Mehl, and R. J. Larsen, "A positivity bias in written and spoken english and its moderation by personality and gender," Social Psychological and Personality Science, vol. 2, no. 5, pp. 508-515, 2011.

[22] D. Garcia, A. Garas, and F. Schweitzer, "Positive words carry less information than negative words," EPJ Data Science: A SpringerOpen Journal, 1(3):1-12, 2012.

[23] T. H. Huang, H. C. Yu, and H. H. Chen, "Modeling pollyanna phenomena in Chinese sentiment analysis," in Proc. 24th International Conf. on Computational Linguistics, Demo. Mumbai, India, 2012, pp. 231-238.

[24] K. Sornig, "Some remarks on linguistic strategies of persuasion," Language, Power and Ideology: Studies in Political Discourse, pp. 95-115, 1989.