Morpheme-based Derivation of

Bipolar Semantic Orientation of Chinese Words

Raymond W.M. Yuen, Terence Y.W. Chan, Tom B.Y. Lai, O.Y. Kwong, Benjamin K.Y. T'sou
Language Information Sciences Research Centre, the City University of Hong Kong
83 Tat Chee Avenue, Hong Kong
{ wmyuen, dcywchan, cttomlai, rlolivia, rlbtsou}@cityu.edu.hk

Abstract

The evaluative character of a word is called its semantic orientation (SO). A positive SO indicates desirability (e.g. Good, Honest) and a negative SO indicates undesirability (e.g., Bad, Ugly). This paper presents a method, based on Turney (2003), for inferring the SO of a word from its statistical association with strongly-polarized words and morphemes in Chinese. It is noted that morphemes are much less numerous than words, and that also a small number of fundamental morphemes may be used in the modified system to great advantage. The algorithm was tested on 1,249 words (604 positive and 645 negative) in a corpus of 34 million words, and was run with 20 and 40 polarized words respectively, giving a high precision (79.96% to 81.05%), but a low recall (45.56% to 59.57%). The algorithm was then run with 20 polarized morphemes, or single characters, in the same corpus, giving a high precision of 80.23% and a high recall of 85.03%. We concluded that morphemes in Chinese, as in any language, constitute a distinct sub-lexical unit which, though small in number, has greater linguistic significance than words, as seen by the significant enhancement of results with a much smaller corpus than that required by Turney.

1. Introduction

The semantic orientation (SO) of a word indicates the direction in which the word deviates from the norm for its semantic group or lexical field (Lehrer, 1974). Words that encode a desirable state (e.g., beautiful) have a positive SO, while words that represent undesirable states (e.g. absurd) have a negative SO (Hatzivassiloglou and Wiebe, 2000). Hatzivassiloglou and Mckeown (1997) used the words 'and', 'or', and 'but' as linguistic cues to extract adjective pairs. Turney (2003) assessed the SO of words using their occurrences near strongly-polarized words like 'excellent' and 'poor' with accuracy from 61% to 82%, subject to corpus size.

Turney's algorithm requires a colossal corpus (hundred billion words) indexed by the AltaVista search engine in his experiment. Undoubtedly, internet texts have formed a very large and easily-accessible corpus. However, Chinese texts in internet are not segmented so it is not cost-effective to use them.

This paper presents a general strategy for inferring SO for Chinese words from their association with some strongly-polarized morphemes. The modified system of using morphemes was proved to be more effective than strongly-polarized words in a much smaller corpus.

Related work and potential applications of SO are discussed in section 2.

Section 3 illustrates one of the methods of Turney's model for inferring SO, namely, Pointwise Mutual Information (PMI), based on the hypothesis that the SO of a word tends to correspond to the SO of its neighbours.

The experiment with polarized words is presented in section 4. The test set includes 1,249 words (604 positive and 645 negative). In a corpus of 34 million word tokens, 410k word types, the algorithm is run with 20 and 40 polarized words, giving a precision of 79.96% and 81.05%, and a recall of 45.56% and 59.57%, respectively.

The system is further modified by using polarized morphemes in section 5. We first evaluate the distinction of Chinese morphemes to justify why the modification can probably give simpler and better results, and then introduce a more scientific selection of polarized morphemes. A high precision of 80.23% and a greatly increased recall of 85.03% are yielded.

In section 6, the algorithm is run with 14, 10 and 6 morphemes, giving a precision of 79.15%, 79.89% and 75.65%, and a recall of 79.50%, 73.26% and 66.29% respectively. It shows that the algorithm can be also effectively run with 6 to 10 polarized morphemes in a smaller corpus.

The conclusion and future work are discussed in section 7.

2. Related Work and Applications

Hatzivassiloglou and Mckeown (1997) presented a method for automatically assigning a + or – orientation label to adjectives known to have some SO by the linguistic constraints on the use of adjectives in conjunctions. For example, 'and' links adjectives that have the same SO, while 'but' links adjectives that have opposite SO. They devised an algorithm based on such constraints to evaluate 1,336 manually-labeled adjectives (657 positive and 679 negative) with 97% accuracy in a corpus of 21 million words.

Turney (2003) introduced a method for automatically inferring the direction and intensity of the SO of a word from its statistical association with a set of positive and negative paradigm words, i.e., strongly-polarized words. The algorithm was evaluated on 3,596 words (1,614 positive and 1,982 negative) including adjectives, adverbs, nouns, and verbs. An accuracy of 82.8% was attained in a corpus of hundred billion words.

SO can be used to classify reviews (e.g., movie reviews) as positive or negative (Turney, 2002), and applied to subjectivity analysis such as recognizing hostile messages, classifying emails, mining reviews (Wiebe et al., 2001). The first step of those applications is to recognize that the text is subjective and then the second step, naturally, is to determine the SO of the subjective text. Also, it can be used to summarize argumentative articles like editorials of news media. A summarization system would benefit from distinguishing sentences intended to present factual materials from those intended to present opinions, since many summaries are meant to include only facts.

3. SO from Association-PMI

Turney (2003) examined SO-PMI (Pointwise Mutual Information) and SO-LSA (Latent Semantic Analysis). SO-PMI will be our focus in the following parts. PMI is defined as:

$$\text{PMI}(word_1, word_2)=\log_2\left(\frac{p(word_1 \,\&\, word_2)}{p(word_1)\,p(word_2)}\right)$$

where p($word_1$ & $word_2$) is the probability that $word_1$ and $word_2$ co-occur. If the words are statistically independent, the probability that they co-occur is given by the product p($word_1$) p($word_2$). The ratio between p($word_1$ & $word_2$) and p($word_1$) p($word_2$) is a measure of the degree of statistical dependence between the words. The SO of a given word is calculated from the strength of its association with a set of positive words, minus the strength of its association with a set of negative words. Thus the SO of a word, *word*, is calculated by SO-PMI as follows:

$$\text{SO-PMI}(word) = \sum_{pword \in Pwords} PMI(word, pword) - \sum_{nword \in Nwords} PMI(word, nword)$$

where Pwords is a set of 7 positive paradigm words (good, nice, excellent, positive, fortunate, correct, and superior) and Nwords is a set of 7 negative paradigm words (bad, nasty, poor, negative, unfortunate, wrong, and inferior). Those 14 words were chosen by intuition and based on opposing pairs (good/bad, excellent/poor, etc.). The words are rather insensitive to context, i.e., 'excellent' is positive in almost all contexts.

A word, *word*, is classified as having a positive SO when SO-PMI(*word*) is positive and a negative SO when SO-PMI(*word*) is negative.

Turney (2003) used the Alta Vista Advanced search engine with a NEAR operator, which constrains the search to documents that contain the words within ten words of one another, in either order. Three corpora were tested. AV-ENG is the largest corpus covering 350 million web pages (English only) indexed by Alta Vista. The medium corpus is a 2% subset of AV-ENG corpus called AV-CA (Canadian domain only). The smallest corpus TASA is about 0.5% of AV-CA and contains various short documents.

One of the lexicons used in Turney's experiment is the GI lexicon (Stone *et al.*, 1966), which consists of 3,596 adjectives, adverbs, nouns, and verbs, 1,614 positive and 1,982 negative.

Table 1 shows the precision of SO-PMI with the GI lexicon in the three corpora.

Percent of full test set	Size of test set	Precision		
		AV-ENG	AV-CA	TASA
100%	3596	82.84%	76.06%	61.26%
75%	2697	90.66%	81.76%	63.92%
50%	1798	95.49%	87.26%	47.33%
25%	899	97.11%	89.88%	68.74%
Approx. no. of words		$1x10^{11}$	$2x10^{9}$	$1x10^{7}$

Table 1: The precision of SO-PMI with the GI lexicon

The strength (absolute value) of the SO was used as a measure of confidence that the words will be correctly classified. Test set words were sorted in descending order of the absolute value of their SO and the top ranked words (the highest confidence words) were then classified. For example, the second row (starting with 75%) in table 1 shows the precision when the top 75% were classified and the last 25% (with lowest confidence) were ignored. We will employ this measure of confidence in the following experiments.

Turney concluded that SO-PMI requires a large corpus (hundred billion words), but it is simple, easy to implement, unsupervised, and it is not restricted to adjectives.

4. Experiment with Chinese Words

In the following experiments, we applied Turney's method to Chinese. The algorithm was run with 20 and then 40 paradigm words for comparison. The experiment details include:

NEAR Operator: it was applied to constrain the search to documents that contain the words within ten words of one another, in either order.

Corpus: the LIVAC synchronous corpus (Tsou *et al.*, 2000, http://www.livac.org) was used. It covers 9-year news reports of Chinese communities including Hong Kong, Beijing and Taiwan, and we used a sub-corpus with about 34 million word tokens and 410k word types.

Test Set Words: a combined set of two dictionaries of polarized words (Guo, 1999, Wang, 2001) was used to evaluate the results. While LIVAC is an enormous Chinese corpus, its size is still far from the hundred-billion-word corpus used by Turney. It is likely that some words in the combined set are not used in the 9-year corpus. To avoid a skewed recall, the number of test set words used in the corpus is given in table 2. In other words, the recall can be calculated by the total number of words used in the corpus, but not by that recorded in the dictionaries. The difference between two numbers is just 100.

Polarity	Total no. of the test set words	Words used in the 9-year corpus
Positive	629	604
Negative	721	645
Total	1350	1249

Table 2: Number of the test set words

Paradigm words: The paradigm words were chosen using intuition and based on opposing pairs, as Turney (2003) did. The first experiment was conducted with 10 positive and 10 negative paradigm words, as follows,

Pwords:誠實(honest), 聰明(clever), 充足(sufficient), 幸運 (lucky), 正確 (right), 優秀 (excellent), 興盛 (prosperous), 善良(kind), 英勇(brave), 謙虛(humble)

Nwords: 虛 僞 (hypocritical), 愚蠢 (foolish), 短缺 (deficient), 不幸(unlucky), 錯誤(wrong), 惡劣(adverse), 衰落(unsuccessful), 殘暴(violent), 懦弱(cowardly), 傲慢(arrogant)

The experiment was then repeated by increasing the number of paradigm words to 40. The paradigm words added are:

Pwords: 溫 和 (mild), 有利 (favourable), 成功 (successful), 正面 (positive), 積極 (active), 樂觀 (optimistic), 良性 (benign), 謹慎 (attentive), 踴躍 (promising), 廉潔(incorrupt)

Nwords: 激進 (radical), 不利 (unfavourable), 失敗 (failed), 負面 (negative), 消極 (passive), 悲觀 (pessimistic), 惡性(malignant), 疏忽(inattentive), 冷淡 (indifferent), 腐敗(corrupt)

4.1 Results

Tables 3 and 4 show the precision and recall of SO-PMI by two sets of paradigm words.

% of test set	100%	75%	50%	25%
Size of test set	1249	937	625	312
Extracted Set	569	427	285	142
Precision	79.96%	86.17%	86.99%	90.16%
Recall	45.56%			

Table 3: Precision and Recall of the SO-PMI of the 20 paradigm word test set

% of test set	100%	75%	50%	25%
Size of test set	1249	937	625	312
Extracted Set	744	558	372	186
Precision	81.05%	86.02%	88.71%	94.09%
Recall	59.57%			

Table 4: Precision and Recall of the SO-PMI of the 40 paradigm word test set

The results of both sets gave a satisfactory precision of 80% even in 100% confidence. However, the recall was just 45.56% under the 20-word condition, and rose to 59.57% under the 40-word condition. The 15% rise was noted.

To further improve the recall performance, we experimented with a modified algorithm based on the distinct features of Chinese morphemes.

5. Experiment with Chinese Morphemes

Taking morphemes to be smallest linguistic meaningful unit, Chinese morphemes are mostly monosyllabic and single characters, although there are some exceptional poly-syllabic morphemes like 葡萄 (grape), 咖啡 (coffee), which are mostly loanwords. In the following discussion, we consider morphemes to be monosyllabic and represented by single characters.

It is observed that many poly-syllabic words with the same SO incorporate a common set of morphemes. The fact suggests the possibility of using paradigm morphemes instead of words.

Unlike English, the constituent morphemes of a Chinese word are often free-standing monosyllabic words. It is note-worthy that words in ancient Chinese were much more mono-morphemic than modern Chinese. The evolution from monosyllabic word to disyllabic word may have its origin in the phonological simplification which has given rise to homophony, and which has affected the efficacy of communication. To compensate for this, many more related disyllabic words have appeared in modern Chinese (Tsou, 1976). There are three basic constructions for deriving disyllabic words in Chinese, including:

(1) combination of synonyms or near synonyms (溫暖, warm, genial, 溫=warm, mild, 暖 =warm, genial)

(2) combination of semantically related morphemes (事情, 事=affair, 情=circumstances)

(3) The affixation of minor suffixes which serve no primary grammatical function (村子, 村 =village, 子=zi, suffix)

The three processes for deriving disyllabic morphemes in Chinese outlined here should be viewed as historical processes. The extent to which such processes may be realized by native speakers to be productive synchronically bears further exploration. Of the three processes, the first two, i.e., synonym and near-synonym compounding, are used frequently by speakers for purposes of disambiguation. In view of this development, the evolution from monosyllabic words in ancient Chinese to disyllabic words in modern Chinese does not change the inherent meaning of the morphemes (words in ancient Chinese) in many cases. The SO of a word often conforms to that of its morphemes.

In English, there are affixal morphemes like dis-, un- (negation prefix), or –less (suffix meaning short-age), -ful (suffix meaning ‘to have a property of’), we can say ‘careful’ or ‘careless’ to expand the meaning of ‘care’. However, it is impossible to construct a word like ‘*ful-care’, ‘*less-care’. However, in Chinese, the position of a morpheme in many disyllabic words is far more flexible in the formation of synonym and near-synonym compound words. For instance, ‘榮’(honor) is a part of two similar word ’榮耀’ (honor-bright) and ‘殊榮’(outstanding-honor). Morphemes in Chinese are like a ‘zipped file’ of the same file types. When it unzips, all the words released have the same SO.

5.1 Probability of Constituent Morphemes of Words with the Same SO

Most morphemes can contribute to positive or negative words, regardless of their inherent meaning. For example, ‘幸’ (luck) has inherently a positive meaning, but it can construct both positive word ‘幸運’ (lucky) or a negative word ‘不幸’ (unlucky). Thus it is not easy to define the paradigm set simply by intuition. But we can assign a probability value for a morpheme in forming polarized words on the basis of corpus data.

The first step is to come up with possible paradigm morphemes by intuition in a large set of polarized words. With the LIVAC synchronous corpus, the types and tokens of the words constructed by the selected morphemes can easily be extracted. The word types, excluding proper nouns, are then manually-labeled as negative, neutral or positive. Then to obtain the probability that a polar morpheme generates words with the same SO, the tokens of the polarized word types carrying the morpheme are divided by the tokens of all word types carrying the morpheme. For example, given a negative morpheme, m_1, the probability that it appears in negative words in token, P(m_1, -ve) is given by:

$$\frac{\text{Tokens of NegativeWo rdtypes Carrying m}_1}{\text{Tokens of All Wordtypes Carrying m}_1}$$

Positive morphemes can be done likewise. Ten negative morphemes and ten positive morphemes were chosen as in table 5. Their values of P(*morpheme*, orientation) are all above 0.95.

	+ve Morpheme	-ve Morpheme
1	獎 (gift)	傷(hurt)
2	勝 (win)	貪(greedy)
3	優 (good)	疑(doubt)
4	堅 (secure)	困(difficult)
5	富 (rich)	急(rush)
6	健 (health)	妄(rash)
7	歡 (happy)	爆(explode)
8	榮 (honor)	禁(ban)
9	努(hardworking)	倒(collapse)
10	順(smooth)	拒(reject)
Derived Types	7383	2048
Tokens	247249	166335

Table 5: Selected positive and negative morphemes

Those morphemes were extracted from a 5-year subset of the LIVAC corpus. A morpheme, free to construct new words, may construct hundreds of words but those words with extremely low frequency can be regarded as ‘noise’. The ‘noise’ may be ‘creative use’ or even incorrect use. Thus, the number of ready-to-label word types formed from a particular morpheme was limited to 50, but it must cover 80% of the tokens of all word types carrying the morpheme in the corpus (i.e., 80% dominance). For example, if the morpheme m_1 constructs 120 word types with 10,000 tokens, and the first 50 high-frequency words can reach 8,000 tokens, then the remaining 70 low-frequency word types, or noise, are discarded. Otherwise, the number of sampled words would be expanded to a number (over 50) fulfilling 80% dominance.

5.2 Results and Evaluation

In table 6, the precision of 80.23% is slightly better than 79.96% of the 20-word condition, and just 1% lower than that of the 40-word condition. However, the recall drastically increases from 45.56%, or 59.57% under the 40-word condition, to 85.03%. In other words, the algorithm run with 20 Chinese paradigm morphemes resulted not only in high precision but also much higher recall than Chinese paradigm words in the same corpus.

% of test set	100%	75%	50%	25%
Size of test set	1249	937	625	312
Extracted Set	1062	797	531	266
Precision	80.23%	85.44%	90.96%	96.61%
Recall	85.03%			

Table 6: Precision and Recall of SO-PMI of the 20 paradigm morpheme test set

Since the morphemes were chosen from a subset of the corpus for evaluation, we repeated the experiment in a separate 1-year corpus (2001-2002). The results in table 7 reflect a similar pattern in the two corpora – both words and morphemes can get high precision, but morphemes can double the recall of words.

	40 Words	20 Morphemes
Size of test set	1065	
Extracted Set	333	671
Precision (Full Set)	75.38%	73.62%
Recall	31.27%	63.00%

Table 7: Precision (full test set only) and Recall of SO-PMI of 40 paradigm words and 20 paradigm morphemes in 1-year corpus

It is assumed that a smaller corpus easily leads to the algorithm's low recall because many low-frequency words in the test set barely associate with the paradigm words. To examine the assumption, the results were further analyzed with the frequency of the test set words. First, the occurrence of the test set words in the 9-year corpus was counted, then the median of the frequency, 44 in this case, was taken. The results were divided into two sections from the median value, and the recall of two sections was calculated respectively, as in table 8.

	Freq≥ Median	Freq< Median
20 Morphemes	99.80%	67.66%
40 Words	89.45%	26.55%

Table 8: Morpheme-based and word-based recall of high-frequency and low-frequency words

The results showed that high-frequency words could be largely extracted by the algorithm with both morphemes (99.80% recall) and words (89.45% recall). However, paradigm words gave 26.55% recall of low-frequency words, whereas paradigm morphemes gave 67.66%. They showed that morphemes outperform words in the retrieval of low-frequency words.

Colossal corpora like Turney's hundred-billion-word corpus can compensate for the low performance of paradigm words in low-frequency words. Such a large corpus has been easily-accessible since the emergence of internet, but it is not cost-effective to use the Chinese texts from the internet because those texts are not segmented. Another way of compensation is the expansion of paradigm words, but doubling the number of paradigm words just raised the recall from 45.56% to 59.57%, as shown in section 4. The supervised cost is not reasonable if the number of paradigm words is further expanded.

Morphemes, or single characters in Chinese, naturally occur more frequently than words in an article, so 20 morphemes can be more discretely-distributed over texts than 20 or even 40 words. The results show that some morphemes always retain their inherent SO when becoming constituents in other derived words. Such morphemes are like a zipped file of the same SO, when the algorithm is run with 20 paradigm morphemes, it is actually run by thousands of paradigm words. Consequently, the recall could double while the high precision was not affected.

It may be argued that the labour cost of defining the SO of 20 morphemes is not sufficiently low either. The following experiments will demonstrate that decreasing the number of morphemes can also give satisfactory results.

6. Experiment with different number of morphemes

The following experiments were done respectively by decreasing the number of morphemes, i.e., 14 and 10 morphemes, chosen from table 5. The algorithm was then run with 3 groups of 6 different morphemes, in which the morphemes were different, and the combination of morphemes in each group was random. The morphemes in each group are shown in table 9. Other conditions for the experiments were unchanged.

6.1 Results and Evaluation

Table 10 shows the results with different number of morphemes, and table 11 shows those for different groups of 6 morphemes. For convenient comparison, the tables only show the results of the full test set, i.e., no threshold filtering.

It is shown that the recall falls as the number of morphemes is reduced. However, even the average recall 66.29% under the 6-morpheme condition is still higher than that under the 40-word condition (59.57%). In section 5, it was evaluated that low recall could be attributed to the low frequency of test set words. Therefore, 6 to 10 morphemes are already ideal for deducing the SO of high-frequency words.

		Number of morphemes used					
	Morpheme	20	14	10	6 (Gp1)	6 (Gp2)	6 (Gp3)
P	獎 (gift)	1			1		
P	優 (good)	1	1	1	1		
P	歡 (happy)	1	1		1		
P	富 (rich)	1	1			1	
P	榮 (honor)	1	1	1		1	
P	順(smooth)	1	1	1		1	
P	勝 (win)	1					1
P	堅 (secure)	1					1
P	健 (health)	1	1	1			1
P	努 (hardworking)	1	1	1			
N	疑(doubt)	1	1	1	1		
N	爆(explode)	1	1		1		
N	禁(ban)	1	1	1	1		
N	妄(rash)	1	1	1		1	
N	貪(greedy)	1	1	1		1	
N	困(difficult)	1	1	1		1	
N	傷(hurt)	1	1				1
N	急(rush)	1					1
N	倒(collapse)	1					1
N	拒(reject)	1					

Table 9: Morphemes selected for different experimental sets, P=+ve, N=-ve, 1='selected', Gp= Group

	Number of morphemes used		
No of morphemes	20	14	10
Size of test set	1249	1249	1249
Extracted Set	1062	993	915
Precision (%)	80.23	79.15	79.89
Recall (%)	85.03	79.50	73.26

Table 10: Precision and Recall of SO-PMI of the test set words with different no. of morphemes

Group of Morphemes	Group 1	Group 2	Group 3	Average
Size of test set	1249	1249	1249	1249
Extracted Set	837	776	871	828
Precision (%)	79.69	78.48	68.77	75.65
Recall (%)	67.01	62.13	69.74	66.29

Table 11: Precision and Recall of SO-PMI of the test set words with 3 different groups of 6 morphemes

The precision remains high from 20 morphemes to 6 morphemes, but from table 10 the precision varies with different sets of morphemes. Group 3 gave the lowest precision of 68.77%, whereas other groups gave a high precision close to 80%. The limited space of this paper cannot allow a detailed investigation into the reasons for this result, only some suggestions can be made.

The precision may be related to the dominant lexical types of the words constructed by the morphemes and those of the test set words. Lexical types should be carefully considered in the algorithm for Chinese because Chinese is an isolating language - no form change. For example, the word '復甦' (recover) can appear in different positions of a sentence, such as the following examples extracted from the corpus:

(1)…美國經濟緩步復甦，最終… (...American economy is gradually recovering…)

(2) …大部分人對經濟復甦轉趨悲觀 。(…most people is now pessimistic about the economy recovery)

(3) …不但阻慢復甦， 也令前景難以估計 。(…decelerates the recovery, but also makes the future unpredictable.)

English allows different forms of 'recovery, like 'recovery', 'recovering', 'recovered' but Chinese does not. Lexical types are thus an important factor for the precision performance. Another way of solving the problems of lexical types is the automatic extraction of meaningful units (Danielsson, 2003). Simply, meaningful units are some frequently-used patterns which consist of two or more words. It is useful to automatically extract the meaningful units with SO in future.

Syntactic markers like negation, and creative uses like ironical expression of adding quotation marks can also affect the precision. Here is an example from the corpus: 「老實商人 」 ('HONEST BUSINESSMAN'). The quotation mark 「」(' ' in English) is to actually express the opposite meaning of words within the mark, i.e., HONEST means DISHONEST in this case. Such markers should further be handled, just as with the use of 'so-called'.

6 Conclusion and Future Work

This paper presents an algorithm based on Turney's model (2003) for inferring SO of Chinese words from their association with strongly-polarized Chinese morphemes. The algorithm was run with 20 and 40 strongly-polarized Chinese words respectively in a corpus of 34 million words, giving a high precision of 79.96% and 81.05%, but a low recall of 45.56% and 59.57%. The algorithm was then run with 20 Chinese polarized morphemes, or single characters, in the same corpus, giving a high precision of 80.23% and an even high recall of 85.03%. The algorithm was further run with just 14, 10 and 6 morphemes, giving a precision of 79.15%, 79.89% and 75.65%, and a recall of 79.50%, 73.26% and 66.29% respectively.

Thus, conveniently defined morphemes in Chinese enhance the effectiveness of the algorithm by simplifying processing and yielding better results even in a smaller corpus compared with what Turney (2003) used. Just 6 to 10 morphemes can give satisfactory results in a smaller corpus.

The efficient application of Turney's algorithm with help of colossal corpus like hundred-billion-word corpus is matched by the ready availability of internet texts. However, the same convenience is not available to Chinese because of the heavy cost of word segmentation.

The efficient application of Turney's algorithm with help of colossal corpus like hundred-billion-word corpus is matched by the ready availability of internet texts. However, the same convenience is not available to Chinese because of the heavy cost of word segmentation.

In our experiment, all syntactic markers are ignored. Better results can be expected if syntactic markers are taken into consideration. An obvious example is negation (not, never) which can counteract the polarity of a word. In future, we will try to handle negation and other syntactic markers.

The lists of the probability of morphemes forming polarized words in section 5.2 can be handled by the concept of decision list (Yarowsky, 2000) which has not been applied in this paper for simplification. In the future, decision lists can be employed to systematically include the loaded features of morphemes.

The experiment can be conducted with different sets of paradigm morphemes, and on corpora of different sizes. With the LIVAC synchronous corpus (Tsou *et al.*, 2000), it should be possible to compare the SO of some words in different communities like Beijing, Hong Kong and Taipei. The data would be valuable for cultural studies if the SO of some words fluctuates in different communities.

SO from association can be also applied to the judgment of news articles like editorials on celebrities. Given a celebrity name or organization name, we can calculate, using SO-PMI, the strength of SO of the 'given word', i.e., the name. Then we would be able to tell whether the news about the target is positive or negative. For example, we tried to calculate the SO-PMI of the name 'George W Bush', the U.S. President, with thousands of polarized Chinese words in the corpus, it was found that the SO-PMI of 'Bush' was about -200 from January to February, 2003, and plunged to -500 from March to April, 2003, when U.S. launched an 'unauthorized war' against Iraq. Such useful applications will be further investigated in future.

References

DANIELSSON, P. 2003. Automatic Extraction of Meaningful Units from Corpora. International Journal of Corpus Linguistics, 8(1), 109-127.

GUO XIAN-ZHEN, ZHANG WEI, LIU JIN, WANG LING-LING. 1999. ChangYong BaoBianYi CiYu XiangJie CiDian (常用褒貶義詞語詳解詞典). Commercial Press, Beijing.

HATZIVASSILOGLOU, V., AND MCKEOWN, K.R. 1997. Predicting the Semantic Orientation of Adjectives. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the ACL, Madrid, Spain, 174-181.

HATZIVASSILOGLOU, V. AND WIEBE, J.M. 2000. Effects of Adjective Orientation and Grad-ability on Sentence Subjectivity. Proceedings of 18th International Conference on Computational Linguistics (Coling'00), Saarbrücken, Germany.

LEHRER, A. 1974. Semantic Fields and Lexical Structure. North Holland, Amsterdam and New York.

STONE, P.J., DUNPHY, D. C., SMITH, M. S., AND OGILVIE, D. M. 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press, Cambridge, MA.

TSOU, B.K. 1976. Homophony and Internal Change in Chinese. Computational Analyses of Asian and African Languages 3, 67-86.

TSOU, B.K., TSOI, W.F., LAI, T.B.Y., HU, J. AND CHAN, S.W.K. 2000. LIVAC, A Chinese Synchronous Corpus, and Some Applications. Proceedings of the ICCLC International Conference on Chinese Language Computing. Chicago, 233-238.

TURNEY, P.D. 2002. Thumbs up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the Association for Computational Linguistics 40th Anniversary Meeting, University of Pennsylvania, Philadelphia, PA, USA.

TURNEY, P.D. & LITTMAN, M.L. 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. ACM Transactions on Information System (TOIS), 21(4), pp315-346.

WANG GUO-ZHANG. 2001. A Dictionary of Chinese Praise and Blame Words (漢語褒貶義詞語用法詞典). Sinolingua, Beijing.

WIEBE, J.M., BRUCE, R., BELL, M. MARTIN, M., AND WILSON, T. 2001. A Corpus Study of Evaluative and Speculative Language. Proceedings of the Second ACL SIG on Dialogue Work-shop on Discourse and Dialogue. Aalborg, Denmark.

YAROWSKY, D. 2000. Hierarchical Decision Lists for Word Sense Disambiguation. Computers and the Humanities, 34(1-2).