Beyond Topicality: Finding Opinionated Chinese Documents

Yejun Wu
School of Library and Information Science, Louisiana State University
267 Coates Hall, Baton Rouge, LA 70803
wuyj@lsu.edu

Douglas W. Oard
College of Information Studies and UMIACS, University of Maryland
Room 4105 Hornbake Bldg (South Wing), College Park, MD 20742
oard@umd.edu

Abstract

The availability of Web 2.0 technologies has made it easy for information users to express their own opinions and access other people's opinions on the Web. We are interested in understanding how opinions expressed in one way by one group compare to opinions expressed in another way by another group, especially in a different language. We have done reasonably well at finding opinionated English mailing lists and blogs, so we started to work on Chinese opinion classification. This paper reports on experiments with a recently released opinion classification test collection for Chinese sentences. Term-scale evidence from a large lexicon and from character-based estimation of semantic orientation for unknown words was used to construct classifiers for subjectivity and polarity that are somewhat more accurate than the best previously reported results. Subjectivity density and the relative predominance of terms with positive and negative semantic orientation were found to be useful features, and appropriate handling of negation was found to be important. With bilingual opinion classification techniques, we can help users find and compare opinions about a topic in two languages.

Introduction

People seek to make sense of their own and relevant others' opinions before they take actions. The availability of Web 2.0 technologies has created an electronic information environment for people to express their own opinions and access others' opinions. Researchers have started to develop techniques to help users make sense of opinionated information in multiple languages.

Needs for opinions

Why do we care about other people's opinions? Many opinions serve to help us understand our world and to make sense of occurrences around us. Opinions represent people's views and feelings about the issues concerned and have implications for potential behavior (Fishbein, 1980). "Our feelings provide us with information. We use our awareness of our feelings to make evaluative judgments and decisions, based on how we feel" (Ekman, 1994, p. 137).

People may have different information seeking tasks concerning opinions. For example, researchers may want to know the aggregate opinions of some population about an issue. Political candidates may wish to know both aggregate opinions regarding their candidacy and which groups of people like/dislike specific positions that they have taken. Policy makers may want to know the attitudes expressed to different audiences by institutional stakeholders (e.g., foreign governments) on an issue. Some searchers want to find a broad range of opinions on a topic; opinion analysis can help us to do a different kind of diversity ranking than the traditional topical diversity (which is also important).

For aggregate presentations of opinions from many documents, personal statements of opinion (such as blogs) are rapidly growing and easily available source of "finger on the pulse" evidence that can augment survey data (which may be higher quality, but available on a far more limited

range of topics, often with relatively long lag times). Aggregating opinions from personal opinion postings may not be better than surveys, but it could be considerably cheaper and thus potentially more easily available because personal opinion postings are available from the Web. With these techniques we can sometimes gain access to more than the opinions of authors. Authors might also report the opinions of others, and blogs and some online news sources (e.g., the Wired Campus (Steele & Wheeler, 2009)) allow comments from readers.

Research on opinion

Social psychologists have been studying opinions and related concepts, such as opinion, emotion, affect, mood, and attitude. We do not discuss their differences here due to space limit. In the library and information science (LIS) community, researchers and practitioners have been doing user studies to understand users' opinions about systems. LIS researchers have started to study the affective issues in information seeking and use (such as emotion, feeling, mood, and sentiment), and an affective paradigm (in contrast to a cognitive paradigm) applied to information behavior in a variety of populations, cultures, and contexts (Nahl & Bilal, 2007). Simply put, social psychologists study the concept of opinion generally, whereas the LIS community has to date focused on user's opinions about sources, systems, and information seeking processes.

In computational linguistics, there has been active research on English opinion classification for more than a decade at the scale of words (Hatzivassiloglou & McKeown, 1997; Turney & Littman, 2003), sentences (Yu & Hatzivassiloglou, 2003; Kim & Hovy, 2004), and entire documents (Pang et al., 2002; Wilson et al., 2005). In contrast to the work in LIS on affective factors, the focus in computational linguistics has been on characterizing opinions about objects and events in the world.

Community efforts to search opinionated English documents have been organized by the Text Retrieval Conference (TREC). In 2005, the TREC Enterprise Search track organized the design of a test collection for identifying emails that contribute at least one statement in favor of or against some specific topic in new (not quoted) text that has been sent by a subscriber to a World Wide Web Consortium (W3C) mailing list. We built one of the best automatic systems (as measured by mean average precision) that was evaluated that year using the W3C test collection (Craswell, de Vries & Soboroff, 2005). In 2006, the TREC Blog track created a new test collection for identifying opinionated blog postings. Again, we built one of the best automatic systems (as measured by mean average precision) that was evaluated that year using the TREC blog test collection (Ounis et al., 2006).

Work on Chinese opinion classification, by contrast, is considerably more recent. Japan's National Institute of Informatics Test Collection for IR systems (NTCIR) evaluation established an Opinion Analysis Pilot Task in 2006, for which they prepared a test collection for Chinese sentence opinion classification.

Purpose of this study

Global networks and the global marketplace for ideas place a premium on understanding how opinions expressed in one way by one group compare to opinions expressed in another way by another group. We are particularly interested in the case in which these groups don't write in the same language. For instance, President Obama's interview with Al Arabiya, an Arabic-language channel based in Dubai, generated subsequent news and blog commentary in different languages around the world (Hutton, 2009). We have chosen the world's two most widely spoken languages, which are the official languages of two of the world's largest economies – English and Chinese – to begin our opinion classification research. Since we have done reasonably well on finding opinionated English mailing lists and blog postings, we focus here on Chinese opinion classification.

In this paper we report on experiments with Chinese sentence opinion classification using the NTCIR-6 test collection for opinion classification. Specifically, we constructed a large Chinese opinion lexicon by leveraging existing Chinese and English lexical resources, we applied character-based classification to estimate the semantic orientation of unknown Chinese words, we tested three ways of handling negation; and we created three aggregate features (subjectivity density, positivity, and negativity) on which we based our classifier designs. The resulting classifiers for opinionated sentences and sentence polarity are somewhat more accurate than the best reported results on the same test collection.

In the next section, we describe the test collection. Subsequent sections then explain how we assembled an opinion lexicon and our process for estimating the semantic orientation of unknown Chinese words and present results from experiments focused on improving subjectivity classification and polarity classification.

Test Collection

The National Taiwan University (NTU) created the NTCIR-6 Opinion Analysis Pilot Task test collection. The collection contains 843 news stories (11,907 sentences) written using traditional Chinese characters that were selected for their relevance to 32 topics using information retrieval techniques. Three native speakers annotated subjectivity for each sentence, and those same annotators recorded polarity judgments for every sentence that they judged as subjective. The agreement (Cohen's kappa) by topic ranges from 0.0537 to 0.4065, with a mean of 0.2328 (Seki et al., 2007), which is considerably lower for this pilot task than would be expected from a more mature annotation process. Our experiment results suggest, however, that the degree of inter-annotator agreement in this first test collection is usually sufficient to reliably distinguish between alternative systems, even with relatively small effect sizes.

NTU combined these judgments to create a single gold standard for subjectivity and polarity by majority voting on a sentence-by-sentence basis. For the Lenient-M standard, two of the three annotators were required to agree on subjectivity. For sentences annotated as subjective at least twice, the polarity [either positive (POS), negative (NEG), or neutral (NEU)] was defined as the polarity with the largest number of votes by the two or three annotators who judged the sentence as subjective. In the case of ties, a reasonable set of decision rules were automatically applied (e.g., one POS and one NEG would result in a Lenient-M annotation of NEU). A Strict-M gold standard was also created in which all three annotators were required to agree on subjectivity, but sparseness made that gold standard less useful for our experiments.

Documents for four topics were released to the NTCIR-6 participants as training data; the remaining 28 topics were used for evaluation. The evaluation set contains 9,240 sentences, of which 5,453 (59.02%) are marked as subjective in Lenient-M. Of those, 2,470 (26.73%) are NEG, 1,209 (13.08%) are NEU, and 1,744 (19.2%) are POS. Therefore, classifiers which always guess the most popular categories (i.e., subjective and negative) would achieve 0.5902 precision and 1.0 recall for subjectivity, and 0.2673 precision and 0.4530 recall for polarity. We show this case as "Baseline B0" in Table 3.

Five research teams participated in the NTCIR-6 Opinion Analysis Pilot Task. In a blind evaluation, the University of Maryland (UMCP) reported the best results for subjectivity (Wu and Oard, 2007) and the Chinese University of Hong Kong (CUHK) reported the best results for polarity (Xu et al., 2007).

Methods

We adopted a shallow linguistic analysis approach to sentence subjectivity and polarity classification in which we rely on evidence about the semantic orientation of each word in the sentence, augmented with appropriate handling for negation.

Lexicon acquisition and preparation

We started with NTU's Chinese opinion lexicon, in which words with positive and negative semantic orientation are annotated (Ku et al., 2006). We augmented this with words for which the semantic orientation had been individually annotated (by NTU) in the (4-topic) training portion of the NTCIR-6 Opinion Analysis Pilot Task test collection. We also manually rekeyed two full books [the "Chinese Positive Dictionary" (Shi & Zhu, 2005) and the "Chinese Negative Dictionary" (Yang & Zhu, 2005)] using traditional Chinese characters (these had originally been published using simplified Chinese characters). Finally, we translated Wilson and Wiebe's English opinion lexicon (Wilson et al., 2005) into Chinese and manually pruned the results.

We created nine mutually exclusive categories: opinion operators (verbs such as *claim* that indicate subjectivity), three categories of opinion words [positive (POS), neutral (NEU), negative (NEG)], opinion indicators (other terms such as *unfortunately* that directly indicate polarity), intensifiers (mainly adverbs such as *very*), quantifiers (mainly subjective quantity terms, such as *a large number of*), negation characters and words (mainly adverbs), and functional negation words (verbs that work semantically as negation, such as *remove*). Four additional categories which overlap POS, NEG and NEU were also created: words that, depending on context, can be both positive and neutral (POS+NEU), negative and neutral (NEG+NEU), positive and negative (POS+NEG), or any of the three (POS+NEU+NEG). The 13 categories are shown in Table 1.

Lexicon translation

An earlier study (Mihalcea et al., 2007) found that translating an English opinion lexicon into Romanian using a bilingual dictionary was useful for Romanian sentence-level subjectivity classification. We acquired Wilson and Wiebe's prior-polarity subjectivity lexicon, which annotates semantic orientation (i.e., POS, NEG, NEU) for 8,221 English words (Wilson et al., 2005). Three English-Chinese lexical resources were used to translate this English lexicon into Chinese: Simpson's English-Chinese dictionary (which used simplified Chinese characters), the ENGLEX English-Simplified Chinese dictionary (automatically converted from simplified to traditional characters, with some errors), and an English-Chinese translation lexicon that was automatically built from parallel text (also automatically converted from simplified to traditional characters, with some errors). This translated lexicon was manually pruned by the first author of this paper to minimize the effect of translation errors, and some category labels were manually corrected to accommodate some of the translation divergence between the two languages. The number of words by category in the resulting lexicon after merging with the Chinese lexical resources described above and deduplication is shown in the "Merge" column of Table 1.

Table 1: Lexicon size.

	NTU	MERGE	EXPAND
POS	2,816	10,937	26,941
NEG	8.276	17,349	43,848
NEU		3,530	9,175
POS+NEU		929	2,736
NEG+NEU		1,281	3,413
POS+NEG		558	1,373
POS+NEU+NEG		263	863
Operator		761	761
Indicator		847	847
Negation		380	380
Functional negation		191	191
Intensifier		302	302
Quantifier		209	209

Lexicon expansion

We can either view Chinese as having many compound words or as combining roots and affixes (Fu, 2006; Yang, 2003). Through introspection, the first author of this paper created a list of Chinese characters and multi-character terms that function as prefixes and suffixes. These were then manually categorized into five rough categories: (1) DE+DI: De as an adjective suffix (such as the English suffix -ive), Di as an adverb suffix (such as -ly); (2) HAVING: verb or adjective prefix, such as You3 (having), Ke2 (can); (3) ABILITY: noun suffix, such as Xing4 (-ability); (4) MAKE: verb prefix, such as Shi3 (verb prefix), and (5) THING: noun suffix: such as Ren2 (person), Dong1 Xi1 (thing). We used these to expand the lexicon by adding or removing these to create additional words, retaining the semantic orientation of the original. As the "Expand" column of Table 1 shows, this substantially increased the coverage of our lexical categories, thus introducing some degree of robustness against inconsequential (and equally valid) differences in Chinese word segmentation.

Estimating semantic orientation of words not in the lexicon

For unknown words, Ku et al. (2006) suggested estimating their semantic orientation based on the Chinese characters the word contains. They defined a character's opinion tendency as the relative rate with which that character is seen in positive and negative words in some training lexicon. The number of characters with positive and negative tendencies then provide evidence for the semantic orientation of a word. Wu and Oard (2007) extended this algorithm by removing words containing negation characters or negation bigrams from the training lexicon, stripping off the negation character(s), and adding the remainder of the word back to the lexicon with the opposite polarity. Since the resulting training lexicon lacks negation characters, negation handling must be included at run time as well (i.e., reversing the sense of the classifier result). The result (appropriately normalized) is a value between -1 (NEG) and 1 (POS) for each unknown word expressing the degree of confidence in the classification.

For negation handling, we examined two types of negation words: natural negation words (or characters), such as "Bu2" (not) in "Bu2 Huai4" (not bad), and functional negation words, such as "Ting2 Zhi3" (stop) in "Ting2 Zhi3 Fan4 Zui4" (stop crimes). We tried three approaches:

• "1-word adjacency": negation reverses the semantic orientation of the immediately following word. An example is "Bu2 Huai4" (not bad).

• "2-word adjacency": negation reverses the semantic orientation of the two immediately following words. An example is "Bu2 Tai4 Huai4" (not too bad).

• "dependency": in which syntactic dependency is used to establish the scope for negation. An example is "Bu2 Shi4 Yu1 Chun3 De Cuo4 Wu4" (not a stupid mistake), where what should be negated is "stupid mistake".

We used the Stanford Parser (Levy & Manning, 2003) to detect syntactic dependency. The first author of this paper manually crafted simple rules to scope the dependency relationship for negation after examination of about 25 of the 3,343 parsed Chinese sentences that contain negation words. For an intrinsic measure of the accuracy of our estimates of semantic orientation for unknown words, we adopted a cross-validation strategy. We first randomly partitioned the NTU (POS and NEG only) lexicon into 10 positive and 10 negative subsets. For our "NTU" condition, we trained a classifier on the first 9 partitions and evaluated classification accuracy using the remaining partition. This process was repeated 10 times, each with a different evaluation partition, reporting the mean classification accuracy across those 10 runs. For our "expanded" condition, we added all the additional POS and NEG words from our expanded lexicon (i.e., those that did not appear anywhere in the much smaller NTU lexicon) to 9-partition training sets, but we still tested on just the remaining NTU partition. As Figure 1 shows, about 90% of all words with a known semantic orientation in the NTU lexicon were classified correctly

with some degree of confidence (i.e., 0 threshold). Requiring increased confidence results in fewer classification decisions and thus lower recall (defined as correct/existing), but higher precision (not shown). As would be expected, the expanded lexicon yields a slightly better chance of correct classification when little evidence is available (at low thresholds), while the smaller but somewhat cleaner NTU lexicon results in somewhat higher recall at high confidence thresholds.

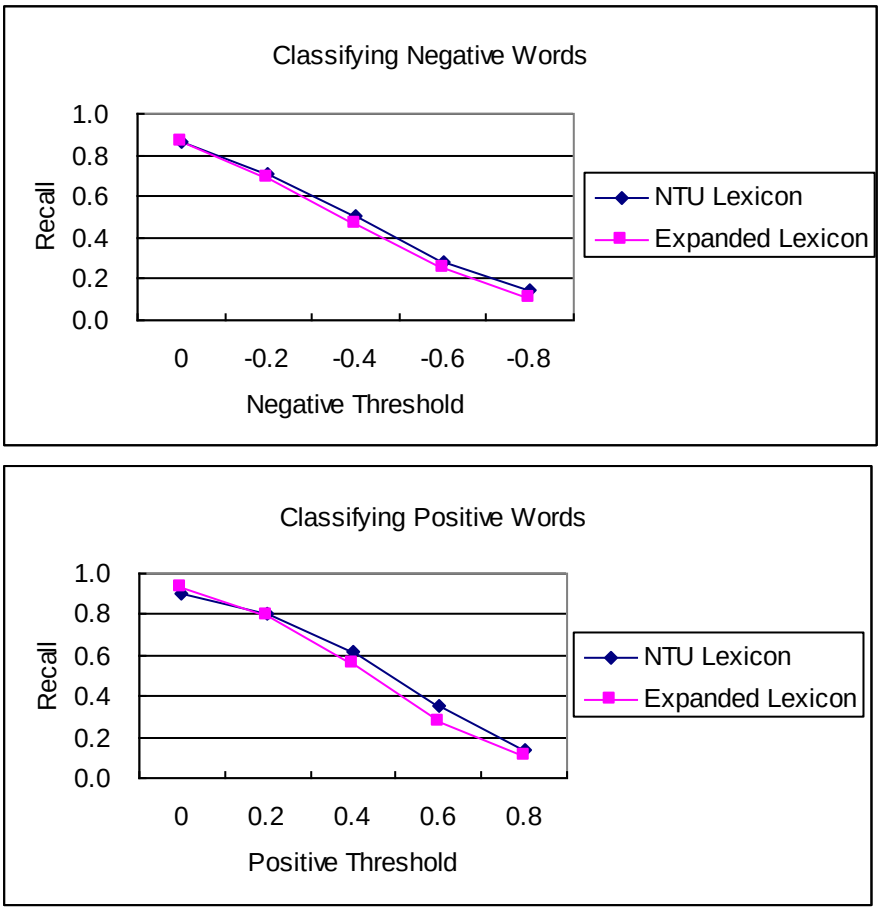

Figure 1: Estimating semantic orientation: effect of confidence threshold on recall.

Classifying Subjectivity and Polarity of Sentences

Once reliable evidence for the semantic orientation of words is available, all that remains to be done is to find the words and combine the resulting evidence. All of our experiments were conducted with the Stanford Segmenter (Tseng et al., 2005). We perform consistent preprocessing for all of our system variants as follows:

(1) Count the number of words in the sentence that can be found in the following parts of the expanded lexicon: POS, NEG, NEU, POS+NEU, NEG+NEU, POS+NEU+NEG, operator, indicator, or intensifier. Then divide that count by the number of words in the sentence to compute subjectivity density.

(2) Working left to right, form running sums of positivity and negativity. For positivity, increment the count when a POS word not preceded by negation (or a NEG word preceded by negation) is encountered. Also increment positivity when a POS+NEU word not preceded by negation (or a NEG+NEU word preceded by negation) is encountered if at that point in the running sums

positivity strictly exceeds negativity. Except where otherwise indicated, all experiments used one-word adjacency negation. At each word, decrement the count for negativity instead if the opposite conditions are satisfied. Note that positivity > 0, while negativity < 0 and neutrality = 0.

(3) If the sentence has at least 1 opinion operator or if its subjectivity density is ≥ 0.5, mark it as *apparently strongly subjective*. Otherwise, mark it as *apparently weakly subjective*. We tuned this parameter 0.5 by hand on the evaluation set.

(4) If the sentence is *apparently weakly subjective* and if positivity and negativity are equal, mark the sentence as *seemingly neutral*.

Improving subjectivity classification

For our B1 baseline system, we classify a sentence as subjective if it is *apparently strongly subjective*. For sentences classified as subjective, we classify their polarity as positive if positivity and negativity sum to a positive number, as negative if they sum to a negative number, and as neutral if they sum to zero. Estimation of semantic orientation for unknown words was not used in this system.

For our A5B1 system, the best of the five variants of B1 that we tried that also lacked estimation of semantic orientation for unknown words, a sentence is classified as subjective if it is *apparently strongly subjective*, or if positivity ≥ 4 or negativity ≤ -4. For sentences classified as subjective, classify their polarity as positive, negative or neutral in the same way as for B1.

Table 2 shows precision, recall, and F_1 for systems B1 and A5B1. B1 detects subjective sentences exactly as well as the best NTCIR-6 system (F=0.7683) we implemented (Seki et al., 2007). Bootstrap resampling confirms that the recall for subjectivity classification is statistically significantly higher for A5B1 than for B1 with the Lenient-M gold standard.

Table 2: Subjectivity-tuned results (Bold: >B1).

Run	Lenient-M	
B0	Subjectivity	Polarity
P	0.5902	0.2637
R	1	0.4530
F	0.7422	0.3362
B1	Subjectivity	Polarity
P	0.7012	0.3328
R	0.8496	0.4033
F	0.7683	0.3649
A5B1	Subjectivity	Polarity
P	0.6941	0.3541
R	0.8790	0.4484
F	**0.7757**	**0.3957**
W3B1	Subjectivity	Polarity
P	0.6953	0.3276
R	0.8740	0.4119
F	**0.7745**	**0.3650**
W6B1	Subjectivity	Polarity
P	Same	0.3330
R	As	0.4148
F	W3B1	**0.3676**
W7B1	Subjectivity	Polarity

Table 4: Polarity-tuned results (Bold: >B1).

Run	Lenient-M	
CUHK	Subjectivity	Polarity
P	0.818	0.522
R	0.519	0.331
F	0.635	0.405
B2	Subjectivity	Polarity
P	0.7012	0.3737
R	0.8469	0.4599
F	0.7683	**0.4095**
C2B2	Subjectivity	Polarity
P	0.6988	0.3728
R	0.8621	0.4599
F	0.7719	**0.4118**
C3B2	Subjectivity	Polarity
P	Same	0.3563
R	as	0.4317
F	B2	0.3904
C6B2	Subjectivity	Polarity
P	0.6985	0.3761
R	0.8617	0.4640
F	0.7716	**0.4154**

P R F	Same As W3B1	0.3066 0.3855 0.3416				

Semantic orientation estimation

We tried eight variants of B1 that leveraged estimation of semantic orientation for unknown words; we report three of those here, as summarized in Table 3.

Table 3: Applying word semantic orientation classification to sentence polarity classification

RUN	NEGATION	THRESHOLD
W3B1	1-word adjacency	$\geq$ 0.8 POS, $\leq$ -0.8 NEG
W6B1	2-word adjacency	$\geq$ 0.8 POS, $\leq$ -0.8 NEG
W7B1	dependency	$\geq$ 0.8 POS, $\leq$ -0.8 NEG

Table 2 includes results for these classifiers as well. For subjectivity detection, bootstrap resampling confirms that the recall of all three variants is statistically significantly higher than that of B1. This indicates that words computed with higher polarity scores have more reliable semantic orientations and are useful for sentence subjectivity detection. We also tried lower thresholds (such as >0 for POS and <0 for NEG), but they proved to be less effective than B1. We have not yet tried W6B1 with A5B1, so we do not know whether the improvements would be cumulative.

For polarity detection, Table 2 shows that W3B1 and W6B1 achieve (statistically significantly) higher recall than B1. Bootstrap resampling identified no significant difference between W3B1 and W6B1, so the apparent slight advantage of 2-word adjacency negation over 1-word adjacency negation is not a reliable indication of an actual difference.

The precision and recall of W7B1 are statistically significantly worse than corresponding values for both B1 and W3B1, indicating that dependency-based negation did not work well. This may result from parse errors, and spot checks of several sentences did find errors. A second potential problem might be the way in which we used the parse results; our simple heuristics might have as-yet undetected errors. So we certainly cannot dismiss the approach on the basis of these experiments alone.

Improving polarity classification

Our best B1 variant (A5B1) represents a slight (1% relative of F1) improvement over the best previously reported results for subjectivity classification, but for polarity classification the best results that we have presented so far were still slightly below CUHK's 0.405 F1 at NTCIR-6. That therefore seemed like a productive direction to focus on next.

When analyzing the results of our B1 variants, we found that there were a larger number of truly negative than truly positive sentences among the *seemingly neutral* sentences. Here, we hypothesize that if a sentence is *apparently strongly subjective* but it has zero aggregated polarity (positivity plus negativity), it would be reasonable to classify it as negative. Therefore we created a B2 baseline that is identical to B1 except that sentences that were classified as neutral by B1 are classified as negative by B2.

As Table 4 shows, that actually works fairly well, yielding an F1 measure slightly better than that reported at CUHK for the Lenient-M standard at NTCIR-6 (Xu et al., 2007). Of course, we have tuned somewhat to this collection, so our goal in making this comparison is just to establish that B2 is a reasonable baseline. Our operating point is, however, clearly quite different from that of the CUHK system, with their system better for precision and our B2 system better for recall. This

suggests that some form of result fusion might be productive, although we have not yet pursued that option. Instead, we took this opportunity to further our investigation of negation handling.

• C2B2: This is essentially a combination of B2 and A5B1 in which we use B2 to compute polarity for the *apparently strongly subjective* sentences, and we use A5B1 to compute polarity for *apparently weakly subjective* sentences. If a sentence is weakly subjective, but has a high positivity or negativity score (positivity ≥ 4 or negativity ≤ -4), if its aggregated polarity score is >4, positive; <-4, negative; or =0, then neutral. One-word adjacency negation was used.

• C3B2: Similar to B2, except that dependency-based negation was used.

• C6B2: Similar to B2 except that 2-word adjacency negation was used.

We do not have the data to perform bootstrap resampling for CUHK's reported results, but bootstrap resampling does confirm that the precision and recall of B2 for polarity classification is statistically significantly higher than those of B1. Moreover, as Table 4 shows, C2B2 is a further improvement for polarity classification over B2, indicating that at least some of the improvement we saw earlier in A5B1 is complementary to the improvements that result from moving from B1 to B2.

In this case, the recall improvement for polarity classification that results from using 2-word adjacency negation (in C6B2) rather than 1-word adjacency negation (in B2) is statistically significant. Dependency-based negation again does poorly, with both precision and recall for C3B2 being statistically significantly below the corresponding values for B2.

Conclusions

Our results indicate that the NTCIR-6 Opinion Analysis Pilot Task has yielded a useful evaluation resource for sentence-scale subjectivity and polarity classification. We have already achieved some modest improvements over the best previously reported results, and the substantial differences in recall and precision between the two best presently available polarity classifiers suggest that additional improvements might be obtained using a suitable result fusion approach or from further tuning of our operating point. It is important to caveat our reported results as suggestive rather than confirmatory, since we have worked with the entire collection while tuning parameters.

Our integration of word-based and character-based evidence yielded small but statistically significant improvements, and proximity was found to yield more useful evidence for negation than our present dependency-based approach. We would be glad to share our relatively large opinion lexicon with other researchers, which may lead to further enrichment of what we hope might ultimately evolve into a standard resource.

Sentence-scale subjectivity classification turns out to be a relatively easy task in the NTCIR-6 Opinion Analysis Pilot Task test collection, in part because simply choosing the most common category (subjective) yields quite a high baseline. Sentence-scale polarity classification is considerably more difficult, however, at least with the relatively shallow lexically-oriented techniques that we have tried. We are therefore interested in looking more deeply into how native speakers of Chinese both express and perceive opinion, with an eye towards possibly identifying some features that are unique to the language that go beyond the simple character-scale semantics that we have explored to date.

Polarity classification remains a challenge. Although our best present polarity classifier does slightly better than the best polarity classifier in NTCIR-6 (i.e., CUHK), and although we are well above the 0.3362 F measure that would result from always guessing the most common category, we have a long way to go before we will be ready to deploy these components in operational systems. However, subjectivity classification alone can be useful in interactive settings where the

system can retrieve (putatively) opinionated documents, and then the user can then make their own assessment of the polarity of the opinions expressed in the retrieved documents. Therefore, being able to classify subjectivity correctly is, to some extent, more important than being able to classify polarity correctly.

References

Aho, A. & Ullman, J. (1972). *The Theory of Parsing, Translation and Compiling*, Prentice-Hall, Englewood Cliffs, NJ.

Cacioppo, J. & Bernston, G. (1994). Relationship between attitudes and evaluative space: a critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin, 115(3)*, 401–423.

Craswell, N., de Vries, A. & Soboroff, I. (2005). Overview of the TREC-2005 Enterprise Track. In *Proceedings of TREC 2005*.

Ekman, P. & Davidson, R. (1994). Afterword: What is the function of emotions? In P. Ekman & R. Davision (Eds.), *The Nature of Emotion*, Oxford: Oxford University Press.

Fishbein, A. (1980). *Understanding Attitude and Predicting Social Behavior*. Englewood Cliffs, N.J.: Prentice-Hall.

Fu, H. (2006). On the definition and scope of modern Chinese affixes—taking five kinds of modern Chinese textbooks as examples. *Journal of Guizhou Educational Institute*, *5*, 2006.

Hatzivassiloglou, V. & McKeown, K. (1997). Predicting the semantic orientation of adjectives. In *ACL-97*, 174–181.

Hutton, A. (2009). Obama addresses Muslim world on Abab Television. *CBS News*. January 27, 2009. http://www.cbsnews.com/blogs/2009/01/27/politics/politicalhotsheet/entry4754691.shtml

Kim, S. & Hovy, E. (2004). Determining the opinion of opinions. In *Coling 2004*.

Ku, L., Liang, Y. & Chen, H. (2006). Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI Spring Symposium Technical Report SS-06-03*, Palo Alto, California, 2006.

Levy, R. & Manning, C. (2003). Is it harder to parse Chinese, or the Chinese treebank? In *ACL 2003.*

Manning, C. & Schutze, H. (2000). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.

Marneffe, M., MacCartney, B. & Manning, C. (2006). Generating typed dependency parsers from phrase structure parses. In *LREC 2006*.

Mihalcea, R., Banea, C. & Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. In *ACL 2007*, Prague, June 23–30, 2007. 976–983.

Moore, D. & McCabe, G. (1989). *Introduction to the Practice of Statistics* (5th Edition). New York: W. H. Freeman. http://www.insightful.com/Hesterberg/bootstrap/

Nahl, D. & Bilal, D. (Eds.) (2007). *Information and Emotion: The Emergent Affective Paradigm in Information Behavior Research and Theory*. ASIST Monograph Series, Information Today.

Ounis, I. et al. (2006). Overview of the TREC-2006 Blog Track.In *Proceedings of TREC 2006.*

Pang, B., Lee, L. & Vaithyanathan, S. (2002). Thumbs up? opinion classification using machine learning techniques. In *EMNLP 2002*, 79-86.

Seki, Y. et al. (2007). Overview of opinion analysis pilot task at NTCIR-6. In *Proceedings of the Sixth NTCIR Workshop*. May 2007, Japan.

Shi, J. & Zhu, Y. (2005). *Bao Yi Ci Ci Dian* (Positive Dictionary). Sichuan Dictionary Press, China.

Steele, P. & Wheeler, B. (2009). Letter to the editor: librarians and tech officials must get along to meet future challenges. *The Wired Campus*, January 8, 2009. http://chronicle.com/wiredcampus/article/3543/letter-to-the-editor-librarians-and-tech-officials-must-get-along-to-meet-future-challenges (accessed on January 20, 2009).

Tseng, H. et al. (2005). A Conditional Random Field Word Segmenter. In *Fourth SIGHAN Workshop on Chinese Language Processing 2005*.

Turney, P. & Littman, M. (2003). Measuring praise and criticism: inference of semantic orientation from association. *ACM TOIS*, *21*, 315–346.

Voorhees, E. (2003). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, *36*, 697–716.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level opinion analysis. In *HLT-EMNLP 2005*.

Wu, Y. & Oard, D. (2007). NTCIR-6 at Maryland: Chinese opinion analysis pilot task. In *Proceedings of the 6th NTCIR Workshop,* May 2007, Tokyo, Japan.

Wu, Y. & Oard, D. (2007). Computing the opinion polarity of Chinese words and sentences. *Technical Report, UMIACS-TR-2007-24.* May 2007. Institute for Advanced Computer Studies, University of Maryland, College Park.

Xu, R., Wong, K. & Xia, Y. (2007). Opinmine -- opinion analysis system by CUHK for NTCIR-6 pilot task. In *Proceedings of the 6th NTCIR Workshop*. May 2007, Japan.

Yang, L. & Zhu, Y. (2005). *Bian Yi Ci Ci Dian* (Negative Dictionary). Sichuan Dictionary Press, Chengdu, Sichuan, China

Yang, X. (2003). Thoughts on roots and affixes. *Chinese Language Learning, 2*, 2003.

Yu, H. & Hatzivassiloglou, V. (2003). Toward answering opinion questions. In *EMNLP-2003*, 129–136.