Sentiment Analysis:
Capturing Favorability Using Natural Language Processing

Tetsuya Nasukawa
IBM Research, Tokyo Research Laboratory
1623-14 Shimotsuruma, Yamato-shi,
Kanagawa-ken, 242-8502, Japan
nasukawa@jp.ibm.com

Jeonghee Yi
IBM Research, Almaden Research Center
650 Harry Rd, San Jose,
CA, 95120, USA
jeonghee@almaden.ibm.com

ABSTRACT
This paper illustrates a sentiment analysis approach to extract sentiments associated with polarities of positive or negative for specific subjects from a document, instead of classifying the whole document into positive or negative.

The essential issues in sentiment analysis are to identify how sentiments are expressed in texts and whether the expressions indicate positive (favorable) or negative (unfavorable) opinions toward the subject. In order to improve the accuracy of the sentiment analysis, it is important to properly identify the semantic relationships between the sentiment expressions and the subject. By applying semantic analysis with a syntactic parser and sentiment lexicon, our prototype system achieved high precision (75-95%, depending on the data) in finding sentiments within Web pages and news articles.

Categories and Subject Descriptors
I.2.7 Natural Language Processing – *Text analysis.*
H.3.1 Content Analysis and Indexing–*Linguistic processing.*

General Terms
Algorithms, Experimentation.

Keywords
sentiment analysis, favorability analysis, text mining, information extraction.

INTRODUCTION
A technique to detect favorable and unfavorable opinions toward specific subjects (such as organizations and their products) within large numbers of documents offers enormous opportunities for various applications. It would provide powerful functionality for competitive analysis, marketing analysis, and detection of unfavorable rumors for risk management.

K-CAP'03, October 23–25, 2003, Sanibel Island, Florida, USA.

For example, enormous sums are being spent on customer satisfaction surveys and their analysis. Yet, the effectiveness of such surveys is usually very limited in spite of the amount of money and effort spent on them, both because of the sample size limitations and the difficulties of making effective questionnaires. Thus there is a natural desire to detect and analyze favorability within online documents such as Web pages, chat rooms, and news articles, instead of making special surveys with questionnaires. Humans can easily recognize natural opinions among such online documents. In addition, it might be crucial to monitor such online documents, since they sometimes influence public opinion, and negative rumors circulating in online documents may cause critical problems for some organizations.

However, analysis of favorable and unfavorable opinions is a task requiring high intelligence and deep understanding of the textual context, drawing on common sense and domain knowledge as well as linguistic knowledge. The interpretation of opinions can be debatable even for humans. For example, when we tried to determine if each specific document was on balance favorable or unfavorable toward a subject after reading an entire group of such documents, we often found it difficult to reach a consensus, even for very small groups of evaluators. Therefore, we focused on finding local statements on sentiments rather than analyzing opinions on overall favorability. The existence of statements expressing sentiments is more reliable compared to the overall opinion. For example,

Product A is good but expensive.

contains two statements. We think it's easy to agree that there is one statement,

Product A is good,

that indicates a favorable sentiment, and there is another statement,

Product A is expensive,

that indicates an unfavorable sentiment. Thus, instead of analyzing the favorability of the whole context, we try to extract each statement on favorability, and present them to the end users so that they can use the results according to their application requirements.

In this paper, we discuss issues of sentiment analysis in consideration of related work and define the scope of our

sentiment analysis in the next section. Then we present our approach, followed by experimental results. We also introduce applications based on our sentiment analysis.

SENTIMENT ANALYSIS

The essential issue in sentiment analysis is to identify how sentiments are expressed in texts and whether the expressions indicate positive (favorable) or negative (unfavorable) opinions toward the subject. Thus, sentiment analysis involves identification of

- Sentiment expressions,
- Polarity and strength of the expressions, and
- Their relationship to the subject.

These elements are interrelated. For example, in the sentence, "*XXX beats YYY*", the expression "*beats*" denotes a positive sentiment toward *XXX* and a negative sentiment toward *YYY*.

However, most of the related work on sentiment analysis to date [1-2,4-5,7,11-14] has focused on identification of sentiment expressions and their polarities. Specifically, the focus items include the following:

- Features of expressions to be used for sentiment analysis such as collocations [12,14] and adjectives [5]
- Acquisition of sentiment expressions and their polarities from supervised corpora, in which favorability in each document is explicitly assigned manually, such as five stars in reviews [2], and unsupervised corpora, such as the WWW [13], in which no clue on sentiment polarity is available except for the textual content [4]

In all of this work, the level of natural language processing (NLP) was shallow. Except for stemming and analysis of part of speech (POS), they simply analyze co-occurrences of expressions within a short distance [7,12] or patterns [1] that are typically used for information extraction [3,10] to analyze the relationships among expressions. Analysis of relationships based on distance obviously has limitations. For example, even when a subject term and a sentiment term are contained in the same sentence and located very close to each other, the subject term and the sentiment term may not be related at all, as in

Although XXX is terrible, YYY is in fact excellent,

where "*YYY*" is not "*terrible*" at all.

One major reason for the lack of focus on relationships between sentiment expressions and subjects may be due to their applications. Many of their applications aim to classify the whole document into positive or negative toward a subject of the document that is specified either explicitly or implicitly [1-2,11-13], and the subject of all of the sentiment expressions are assumed to be the same as the document subject. For example, the classification of a movie review into positive or negative [2,13] assumes that all sentiment expressions in the review represent sentiments directly toward that movie, and expressions that violate this assumption (such as a negative comment about an actor even though the movie as a whole is considered to be excellent) confuse the judgment of the classification. On the contrary, by analyzing the relationships between sentiment expressions and subjects, we can make in-depth analyses on what is favored and what is not.

In this paper, we define the task of our sentiment analysis as to *find sentiment expressions for a given subject and determine the polarity of the sentiments.* In other words, it is to identify text fragments that denote a sentiment about a subject within documents rather than classifying each document as positive or negative towards the subject. In this task, the identification of semantic relationships between subjects and sentiment-related expressions is a key issue because the polarity of the sentiment may be entirely different depending on the relationships, as in the above example of "*XXX beats YYY.*" In our current implementation, we manually built the sentiment lexicon based on the requirements discussed in the next section.

FRAMEWORK OF SENTIMENT ANALYSIS

Definition of Sentiment Expressions

Besides adjectives, other content words such as nouns, adverbs, and verbs are also used to express sentiments. In principle, a sentiment expression using an adjective, say "*good*", denotes the sentiment towards its modifiee noun such as in "*good product*," and the whole noun phrase ("*good product*") itself becomes a sentiment expression with the same polarity as the sentiment adjective (positive for "*good*" in this case). Likewise, a sentiment expression using an adverb, say "*beautifully*," denotes the sentiment towards its modifiee verb such as in "*play beautifully*," and the polarity of the sentiment is inherited by the modifiee verb. Thus, sentiment expressions using adjectives, adverbs, and nouns can be simply defined as either positive or negative in terms of polarity. In contrast, as in the examples in the previous section such as "*XXX beats YYY*," the polarity of sentiments denoted by the sentiment expressions in verbs may depend on the relationships with their arguments. In this case, positive sentiment is directed towards its subject and negative sentiment is directed towards its object. In addition, some verbs do not denote sentiment by themselves, but only transfer sentiments among their arguments. For example, a be-verb transmits the sentiment of its complement to its subject such as in "*XXX is good*," in which the positive sentiment of its complement, "*good*," is transferred to its subject, "*XXX*." Thus, we classified sentiment-related verbs into two types, specifically,

- Sentiment verbs that direct either positive or negative sentiment toward their arguments,

- Sentiment transfer verbs that transmit sentiments among their arguments,

and associate them with arguments such as subjects and objects that inherit or provide sentiment.

Therefore, we have manually defined sentiment expressions in a sentiment lexicon by using a simple notation that consists of the following information:

- Polarity
  positive (good), negative (bad), or neutral is denoted by **g**, **b**, or **n**, respectively, and sentiment transfer verbs are denoted by **t**.
- Part of speech (POS)
  Currently, adjective (**JJ**), adverb (**RB**), noun (**NN**), and verb (**VB**) are registered in our lexicon
- Sentiment term in canonical form
- Arguments such as subject (**sub**) and object (**obj**) that receive sentiment from a sentiment verb or arguments that provide sentiment to and receive sentiment from a sentiment transfer verb

For example, the following notation

gVB admire obj

indicates that the verb "*admire*" is a sentiment term that indicates favorability towards a noun phrase in its object when the noun phrase in the object contains a subject term. Likewise,

bVB accuse obj

indicates that the verb "*accuse*" is a sentiment term that indicates unfavorability against a noun phrase in its object when the noun phrase contains a subject term.

bVB fail sub

indicates that the verb "*fail*" is a sentiment term that conveys unfavorability towards a noun phrase in its subject when the noun phrase contains a target subject term.

tVB provide obj sub

indicates that verb "*provide*" passes the (un)favorability of its object into its target subject term if the object noun phrase contains (un)favorability and the target term is in its subject, such as in,

"*XXX provides a good working environment*."

"*XXX provides a bad working environment*."

where "*XXX*" is a subject term with favorable and unfavorable sentiment, provided that "*a good working environment*" and "*a bad working environment*" are favorable and unfavorable, respectively.

Finally,

tVB prevent obj ~sub

indicates that the verb "*prevent*" passes the opposite of the (un)favorability of its object to its target subject term if the object noun phrase contains (un)favorability and the target term is in its subject, such as in,

"*XXX prevents trouble*."

in which "*XXX*" is a subject term receiving favorable sentiment, and "*trouble*" is a sentiment term for unfavorability.

For terms with other POS, we simply classify them into favorable, unfavorable, and neutral. For example,

bJJ crude

indicates the adjective (denoted by JJ) "*crude*" has unfavorable sentiment (denoted by "b" in the first column), and

nNN crude oil

indicates that the noun phrase (denoted by NN) "*crude oil*" is neutral (denoted by "n" in the first column) so that the term "*crude*" in "*crude oil*" is not treated as a negative sentiment. Thus, sentiment terms can be compound words, and they are applied using the leftmost longest match method so that longer terms with more matching elements are favored. In addition, we also allowed the use of regular expressions for the flexibility of expressions such as

bVB put \S+ at risk sub,

in which "\S+" can be matched with one or more sequences of non-whitespace characters, and a sentence such as

"*XXX put communities at risk*."

is considered to be negative for XXX.

In principle, we tried to define the framework of the sentiment lexicon as simply as possible, both to ease the manual work and for the sake of simplifying automatic generation in the future. As we deal with natural language, we may find exceptional cases in which sentiments defined in the lexicon do not hold. For example, "*put something at risk*" may be favorable when the "*something*" is unfavorable such as the case of "*hackers*." Thus, we started with basic entries that cover most of the cases properly and dealt with exceptional cases by adding entries that deal with more specific terms to be applied properly in those specific cases.

Currently, we have 3,513 entries in the sentiment analysis dictionary, as summarized in Table 1. Among these entries, regular expressions were used in 14 cases.

Table 1. Distribution of sentiment terms

POS	Total	positive	negative	neutral
adjective	2,465	969	1,495	1
adverb	6	1	4	1
noun	576	179	388	9
Sentiment verb	357	103	252	2
Transfer verb	109			

Algorithm

We applied sentiment analysis to text fragments that consist of a sentence containing a subject term and the rest of the following paragraph. The window always included at least 5 words before and 5 words after the target subject. There is an upper limit of 50 words before and 50 words after. Thus, the task of our sentiment analysis approach is to find sentiment expressions that are semantically related to the subject term within the text fragment, and the polarity of the sentiment. The size of this text fragment was defined tentatively based on our preliminary analysis to capture the minimal required context around the subject term.

In order to identify sentiment expressions and analyze their semantic relationships with the subject term, natural language processing plays an important role. POS tagging allows us to disambiguate some polysemous expressions such as "*like*," which denotes sentiment only when used as a verb instead of as an adjective or preposition. Syntactic parsing allows us to identify relationships between sentiment expressions and the subject term. Furthermore, in order to maintain robustness for noisy texts from various sources such as the WWW, we decided to use a shallow parsing framework that identifies phrase boundaries and their local dependencies in addition to POS tagging, instead of using a full parser that tries to identify the complete dependency structure among all of the terms.

For POS tagging, we used a Markov-model-based tagger essentially the same as the one described in [6]. This tagger assigns a part of speech to text tokens based on the distribution probabilities of candidate POS labels for each word and the probability of a POS transition extracted from a training corpus. We used a manually annotated corpus of Wall Street Journal articles from the Penn Treebank Project [9] as the training corpus. For these experiments, the tagger was configured to treat unknown words (i.e. those not seen in the training corpus, and excluding numbers) as nouns. The tagger uses a lexical look-up component, which offers sophisticated inflectional analysis for all known words.

After a POS for each word was assigned, we used shallow parsing in order to identify phrase boundaries and local dependencies, typically binding subjects and objects to predicates. This shallow parsing is based on the application of a cascaded set of rules, successively identifying more and more complex phrasal groups. Thus simple patterns can find simple noun groups and verb groups, and these can be composed into a variety of complex NP configurations. At a yet higher level, clause boundaries can be marked, and even (nominal) arguments for (verb) predicates can be identified. These POS tagging and shallow parsing functionalities have been implemented using the Talent System based on the TEXTRACT architecture [8].

After obtaining the results of the shallow parser, we analyze the syntactic dependencies among the phrases and look for phrases with a sentiment term that modifies or is modified by a subject term. When the sentiment term is a verb, we identify the sentiment according to its definition in the sentiment dictionary. Syntactic subjects in passive sentences are treated as objects for matching argument information in the definition. Finally, a sentiment polarity of either +1 (positive = favorable) or -1 (negative = unfavorable) is assigned to the sentiment according to the definition in the dictionary unless negative expressions such as "*not*" or "*never*" are associated with the sentiment expressions. When the negative expressions are associated, we reverse the polarity. As a result,

- The polarity of the sentiments,
- The sentiment expressions that are applied, and
- The phrases that contain the sentiment expressions,

are identified for a given subject term.

The following examples were output from our current prototype system, as applied to genuine texts from the WWW. In each input, we underlined the subject term that our system targeted for analysis. Each output starts with an indicator of sentiment polarity toward the subject. The subject term and sentiment terms identified in the input are connected with "---" with their representation in canonical forms that are associated with the whole phrase in the parenthesis that contains them. When transfer verbs are used, information on the transfer verbs appears in the middle of the representation of the subject term and sentiment term. Among the following examples, Example 3 contains negation, and Example 4 is a passive sentence. All of the typographic errors in the following examples, including the ones in the next section, came from the original texts, and similar problems were usually handled properly by our shallow parser.

Example 1:

<input> (subject="MDX")

For 2002, the MDX features the same comfort and exhilaration, with an even quieter ride.

<output>

+1 MDX (the MDX)---feature (features)---comfort (the same comfort and exhilaration)

Example 2:

<input> (subject="IBM")

Of the seven stakeholder groups, IBM received the highest score in the ranking for its policies and programs for minorities and women.

<output>

+1 IBM (IBM)---receive (received)---high score (the highest score in the ranking)

Example 3:

<input> (subject="canon")

Image quality was 1 and the Canon G2 definately did not disappoint me! (sic.)

<output>

+1 canon (the Canon G2 definately)---disappoint (did not disappoint)

Example 4:

<input> (subject="Range Rover")

They haven't, even though the Range Rover was celebrated as a status symbol as long ago as the 1992 movie The Player.

<output>

+1 celebrate (was celebrated)---Range Rover (SUB the Range Rover)

Example 5:

<input> (subject="Ford Explorer")

For example, the popular Ford Explorer retains about 75 percent of its sticker price after three years, while the high-end Lincoln Continental retains only about half of its original cost after the same amount of time.

<output>

+1 popular---Ford Explorer (the popular Ford Explorer)

EXPERIMENTAL RESULTS

We have applied this sentiment analysis method to data in a number of domains, and evaluated the results manually by using a benchmark corpus and other open test data. For the evaluations, we checked if the polarity of the sentiment was appropriately assigned to the given subject in each input in terms of the sentiment expression in the output, and calculated the precision and recall. Precision is the ratio of correct cases within the system outputs. Recall is the ratio of correct cases that the system assigned compared to the base of all cases where a human analyst associated either positive or negative sentiments manually. In other words, precision and recall are calculated with the following formulas:

A = number of all cases that the system assigned either a positive or negative sentiment

B = number of all cases that the human assigned either a positive or negative sentiment

C = number of correct cases in the system output based on the manual judgment

Precision = C/A

Recall = C/B

Evaluation with Benchmark Corpus

In order to evaluate the quality of the sentiment analysis, we created a benchmark corpus that consists of 175 cases of subject terms within contexts extracted from Web pages from various domains. Each case was manually identified to represent either a favorable or an unfavorable sentiment toward the subject. There were 118 favorable cases and 58 unfavorable cases. The examples in the previous section were taken from this corpus.

After modifying the dictionary for the benchmark corpus by adding appropriate terms, our current prototype system achieved 94.3% precision and 28.6% recall as it extracted sentiments for 53 cases (50 correct).

Evaluation with Open Test Corpus

In order to verify the quality for practical use, we used the prototype for a new test set with 2,000 cases related to camera reviews, also from Web pages. This time, about half of the cases contained either favorable or unfavorable sentiments and the other half were neutral. Our system extracted sentiments for 255 cases, and 241 of them were correct in terms of the polarity of either negative or positive toward its subject within the context. Thus, without any modification of the dictionary, the current prototype system achieved 94.5% (=241/255) precision with about 24% (=241/1,000) recall.

Analysis of Failures

In the open test corpus of camera reviews, our system failed to judge the correct sentiment in cases similar to the following:

Example 6:

<input> (subject="picture")

It's difficult to take a bad picture with this camera.

<output>

-1 bad---picture (a bad picture)

This is a positive statement for the camera, and it's not relevant to extract this "*bad picture*" as a negative sentiment.

Example 7:

<input> (subject="canon")

The g2 is the daddy of all slr type cams for users that dont make their money of photographing and probably a good choise for them who do to all tests ive done and seen shows that the Canon cameras are the best as objection to one of the negative reviews saying canon sucks In my oppinion it beats all fuji nikon minolta sony and other brand competitors. (sic.)

<output>

-1 canon (canon)---suck (sucks)

This may be considered as correct in a sense that it indicates existence of some negative reviews. However, the

whole context is positive toward "*canon*", and we don't think this sentiment is what the author intended, so we counted it as incorrect. In this example, our system also generated the following output for the previous "*Canon*" within the same sentence.

+1 canon (the Canon cameras)---be (are)---best (the best)

Example 8:

<input> (subject="battery")

Also the battery went dead while at Animal Kingdom and one feature I used to like about the Olympus is that if the recharge-able batteries went dead you could just pop some AA's in and still get your pictures.

<output>

-1 battery (the battery)---go (went)---dead (dead)

Here the incident that "*the battery went dead*" is described as a normal event instead of product failure.

As seen in Examples 6 through 8, most of the failures are due to the complex structures of the sentences in the input context that negates the local sentiment for the whole, and they are not due to failures of our syntactic parser. Thus, in order to improve precision, we can restrict the output of ambiguous cases that tend to be negated by predicates at higher levels. For example, sentiments in noun phrases (NPs) as in Examples 5 and 6 can easily be negated by the predicates that they are attached to, so we might consider suppressing the extraction of NP-type sentiments. In addition, sentiments in a sentence that contains an if-clause, as in the following example, are highly ambiguous, as are the sentiments in interrogative sentences.

Example 9:

<input> (subject="AALIYAH")

If AALIYAH was so good, why she is grammyless. Do you like her? Do they know it? Do you like them?

<output>

+1 AALIYAH (AALIYAH)---be (was so)---good (good)

Thus, by suppressing the output of ambiguous sentiments, we can improve the precision fairly easily. In fact, we have observed that we could achieve 99% precision in a data set in the pharmaceutical domain by adding enough entries and eliminating the ambiguous sentiments of the NP-type, since most of the failures were NP-type cases in that data. However, improvement in precision damages recall and it is also important to improve the recall as well as the precision by handling such ambiguous cases properly. By eliminating the ambiguous sentiments, in the benchmark corpus, the precision was improved from 94.3% to 95.5%, but the recall was reduced from 28.6% to 24%, as it extracted sentiments for 44 cases (42 correct) in comparison to 53 cases (50 correct) with the ambiguous ones.

In order to investigate the possibility of improving the recall, we analyzed 122 cases in the benchmark corpus for which our system failed to extract any sentiments. In 14 (11.5%) of these cases, the subject terms and the sentiment expressions did not appear in the same sentence. Anaphora resolution may solve half of these 14 cases by associating anaphoric expressions such as pronouns with their subject terms, since the anaphoric expressions appeared in the same sentences with the sentiment expressions. In the remaining 108 (88.5%) cases, the subject terms and sentiment expressions appeared in the same sentence. In most of these cases, the sentences were quite long and contained nested sub-clauses, embedded sentences and phrases, or complex parallel structures. Since it is quite difficult for a shallow parser to make appropriate analyses for such cases, the failures in these cases are due to either limitations or failures of the shallow parser.

As in the real examples such as Example 7, there are quite a few typographic errors and ill-formed sentences in the Web pages. Thus, in order to maintain robustness for those cases, we decided to continue using a shallow parser instead of a full parser. Yet based on the result that failures in syntactic analysis did not damage the precision, it might make sense to adopt a full parser and make deeper NLP analysis, such as anaphora resolution, in order to improve the recall for longer and more complicated sentences.

APPLICATIONS

In evaluating our system with real-world applications, we have applied it to about a half million Web pages and a quarter million news articles.

First, we extracted sentiments on an organization by defining thirteen subject terms that typically represent the organization, including its full name, its short name, former names, and its divisional names. Out of 552,586 Web pages, 6,415 pages were classified as mentioning the organization after checking for other terms in the same pages. These 6,415 pages contained 16,862 subject references (2.6 references per page). Among them, 369 references were associated with either positive or negative sentiments by our prototype, and the precision was 86%. We also scanned for the same organization in 230,079 news articles. Among these, 1,618 articles were classified as mentioning the organization, and they contained 5,600 subject references (3.5 references per article). A total of 142 references were associated with either positive or negative sentiments, and 88% of them were correct in terms of precision.

We also extracted sentiments about product names. This time, we chose a pharmaceutical domain, and the subjects were the names of ten medicines. Out of 476,126 Web pages, 1,198 pages were classified as mentioning one of the medicines, and there were 3,804 subject references (3.2 references per page). Our prototype system associated 103 references with either positive or negative sentiments, and 91% of them were correct in terms of precision.

Based on these results, we feel that our approach allows us to collect meaningful sentiments from billions of Web pages with relatively high precision. In the following subsections, we introduce two typical applications that can take advantage of our sentiment analysis approach in spite of its relatively low recall, and we discuss important issues for these applications.

Capturing Trends on Sentiments

By comparing the sentiments on specific subjects between uniform intervals we can detect opinion trends. By comparing sentiments for specific subjects with other subjects, we can do a competitive analysis. For example, we can do a quantitative analysis by counting the numbers of positive and negative sentiments to see if a subject is on balance favorable or unfavorable. It may be useful to analyze changes in the balance over some period of time and to compare it with other subjects. The output of our method also allows us to do qualitative analysis easily because it provides very short summaries of the sentiment expressions. For such applications, precision in the polarity is considered to be more important than recall so that users don't have to verify the results by reading the original documents.

In order to verify the credibility of trends detected by the system output in spite of its low recall, we compared the ratio of favorability in the detected sentiments with the missed sentiments by using the data on camera reviews from Web pages. We asked a human evaluator to pick up positive and negative sentiments for brands *A*, *B*, *C*, and *D* from 20,000 text fragments within the open test corpus. As shown in Table 2, the ratio of favorability in system output was comparable to the human evaluation, although we need to conduct larger scale experiments to confirm its statistical significance.

Table 2. Comparison of number of sentiments on camera brands detected by human and system

	polarity	**brand *A***	**brand *B***	**brand *C***	**brand *D***
Human	favor.	437	169	80	39
	unfav.	70	65	51	41
System	favor.	52	22	9	3
	unfav.	4	5	2	1

Finding important documents to be monitored

For some areas of analysis where data tends to be sparse, it is difficult to find relevant documents, and human analysts are willing to read the content of the document that the sentiment analysis approach identified as having sentiments. For example, opinions on corporate images are generally harder to find compared to opinions on products (whose comparisons may be found on various consumer Web sites), and analysts of corporate images may want to read through the relevant Web pages.

For this type of application, recall is more important than the precision in the polarity, and a recall around 20% for finding these documents may be too low. However, according to our experience, a document that contains a sentiment expression usually contains quite a few sentiments, as they express multiple sentiments from various viewpoints or for various subjects to make comparison. Thus, even though the recall of finding a particular sentiment using our approach is around 20% or less, the chances of finding important documents tend to be high enough.

CONCLUSION AND FUTURE WORK

We have illustrated a sentiment analysis approach for extracting sentiments associated with polarity of positive or negative for specific subjects from a document, instead of classifying the whole document as positive or negative. In order to achieve high precision, we focused on identifying semantic relationships between sentiment expressions and subject terms. Since sentiments can be expressed with various expressions including indirect expressions that require common sense reasoning to be recognized as a sentiment, it's been a challenge to demonstrate the feasibility of our simple framework of sentiment analysis. Yet our experimental results indicate we can actually extract useful information on sentiments from most of the texts with our current implementation.

The initial experiments resulted in about 95% precision and roughly 20% recall. However, as we expand the domains and data types, we are observing some difficult data for which the precision may go down to about 75%. Interestingly, that data usually contains well-written texts such as news articles and descriptions in some official organizational Web pages. Since those texts often contain long and complex sentences, our simple framework finds them difficult to deal with.

As seen in the examples, most of the failures are due to the complex structures of sentences in the input context that negates the local sentiment for the whole, and they are not due to failures of our syntactic parser. For example, a complex sentence such as "*It's not that it's a bad camera*" confuses our method. It is noteworthy that failures in parsing sentences do not damage the precision in our approach. In addition, it allows us to classify ambiguous cases by identifying features in sentences such as inclusion of if-clauses and the interrogatives. Thus, we can maximize the precision by eliminating such ambiguous cases for applications that prefer precision rather than recall.

Because of our focus on precision, the recall of our approach remains low. However, it's still effective for various applications. Trend analysis and important document identification in terms of sentiments are typical examples that can take advantage of our approach.

Our current system requires manual development of sentiment lexicons, and we need to modify and add

sentiment terms for new domains. Although our current domain-dependent dictionaries remain relatively small, with fewer than 100 entries each for five different domains, dictionary maintenance would be an important issue for large-scale applications. Thus, we are working toward automated generation of the sentiment lexicons in order to reduce human intervention in dictionary maintenance, both for improving precision for new domains as well as for improving the overall recall.

In addition, for improvement of both precision and recall, we are exploring the feasibility of integrating a full parser and various discourse processing methods including anaphora resolution.

ACKNOWLEDGMENTS

We would like to thank Wayne Nieblack, Koichi Takeda, and Hideo Watanabe for overall support of this work, Roy Byrd, Mary Neff, Bran Bograev, Herb Chong, and Jim Cooper for the use of their POS tagger and shallow parser as well as its Java interface, and Jasmine Novak, Zengyan Zhang, and David Smith for their collaboration and advice on this work. We would also like to thank the anonymous reviewers for their comments and suggestions, and Shannon Jacobs for help in proofreading early versions of this paper.

REFERENCES

[1] Chinatsu Aone, Mila Ramos-Santacruz, and William J. Niehaus. AssentorR: An NLP-Based Solution to E-mail Monitoring. In *Proceedings of AAAI/IAAI 2000*, pages 945-950. 2000.

[2] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP), pages 79-86. 2002.

[3] Ralph Grishman and Beth Sundheim. Message understanding conference - 6: A brief history. In *Proceedings of the 16th International Conference on Computational Linguistics* (COLING), pages 466-471. 1996.

[4] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL*, pages 174-181. 1997.

[5] Vasileios Hatzivassiloglou and Janyce M. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of 18th International Conference on Computational Linguistics* (COLING), pages 299-305. 2000.

[6] Chris Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA. 1999.

[7] Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi, Toshikazu Fukushima. Mining Product Reputations on the Web. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD), pages 341-349. 2002.

[8] Mary S. Neff, Roy J. Byrd, and Branimir K. Boguraev. The Talent System: TEXTRACT Architecture and Data Model. In *Proceedings of the HLT-NAACL 2003 Workshop on Software Engineering and Architecture of Language Technology systems* (SEALTS), pages 1-8. 2003.

[9] Penn Treebank Project. http://www.cis.upenn.edu/treebank/

[10] SAIC Information Extraction. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

[11] Ellen Spertus. Smokey: Automatic recognition of hostile messages. In *Proceedings of the Conference on Innovative Applications of Artificial Intelligence* (IAAI), pages 1058-1065. 1997

[12] Richard M. Tong. An operational system for detecting and tracking opinions in on-line discussions. *Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification*, pages 1-6. 2001.

[13] Peter Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (ACL), pages 417-424, 2002.

[14] Janyce M. Wiebe, Theresa Wilson, and Matthew Bell. Identifying collocations for recognizing opinions. In *Proceedings of the ACL/EACL Workshop on Collocation*. 2001.

[15] Jeonghee Yi and Tetsuya Nasukawa. Sentiment Analyzer: Extracting Sentiments towards a Given Topic using Natural Language Processing Techniques. In *Proceedings of the Third IEEE International Conference on Data Mining* (ICDM). (To appear). 2003.