Question Answering in the Medical Domain: the Role of Clinical Outcome and Polarity

Yun Niu, Xiaodan Zhu and **Graeme Hirst**
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada M5S 3G4
yun,xzhu,gh@cs.toronto.edu

Abstract

Answers to clinical questions are often complicated. In this paper, we formulate the problem as a multi-document summarization task, and construct extraction-based summaries to answer questions about effects of applying a medication to a disease. The experimental results show that identifying clinical outcomes and detecting their polarity improves the performance of summarization. Domain knowledge and context information is effective in obtaining the information about clinical outcomes.

1 Introduction

Medical literature is an important source to help clinicians in patient treatment (Straus and Sackett, 1999). Clinicians often need to consult literature on the latest information in patient care, such as side effects of a medication or symptoms of a disease. As discussed in (Author1 et al., 2003), there are usually many aspects related to a clinical question, and it should not be answered with just one word or phrase. For example:[1]

> **Q:** In a patient with a generalized anxiety disorder, does cognitive behaviour or relaxation therapy decrease symptoms?

Clinical outcomes of cognitive behaviour or relaxation therapy could be complicated. They could be beneficial or harmful; they could have different effects for different patient groups; some clinical trials may show they are beneficial while others don't. Thus, answers to these questions can only be obtained by taking into account various experimental results in the medical research literature and extracting important points from them. This can be formulated as a multi-document summarization task. The question-answering (QA) problem is then to summarize relevant information from high-quality clinical evidence. In this paper, we focus on extracting answers to questions asking about effects of applying a medication to a disease.

Identifying clinical outcomes is likely to help, as we expect that clinical outcomes are important information in this task. As mentioned above, some outcomes may be beneficial, some may be harmful or neutral. Since contradictory evidence can be crucial in answering a clinical question, it may be helpful to detect the polarity of a clinical outcome. We notice that a single sentence usually describes one or more clinical outcomes completely. Starting with that observation, our work is to construct sentence-level extracts as answers using information of clinical outcomes and their polarity. To our knowledge, no work has been published on a similar task. Yu and Hatzivassiloglou (2003) detects sentence-level opinions and their semantic orientation in news articles. Although they mentioned that polarity information is applied in their QA system, no details about how it is incorporated is described in the paper, nor evaluation of the effectiveness of this information. In contrast, it is easy to see that outcome identification and polarity detection has important and direct connection to our task.

[1] This example is taken from a collection of questions that arose over a two-week period in August 2001 in a clinical teaching unit at the University of ***.

2 *Clinical Evidence* as a benchmark

Evaluation of a multi-document summarization system is difficult, especially in the medical domain where there is no standard annotated corpus available. However, we note that the book *Clinical Evidence* (CE) (Barton, 2002) provides a standard to evaluate our work against. CE is a publication that reviews and consolidates experimental results for clinical problems; it is updated every six months. Each section in CE is about a clinical problem (disease). A section is divided into several subsections. Each subsection summarizes the evidence concerning a particular medication (or a class of medications) for the problem, including results of clinical trials on the benefits and harms of the medication. The information sources that CE reviews include medical journal paper abstracts, review articles, textbooks, etc. Human experts work on the collected information and summarize them to get concise evidence on every specific topic. This is the process of multi-document summarization. Thus, evidence in subsections of CE can be regarded as human-written multi-document summaries of the papers (abstracts) that are cited in that subsection. Reference to each piece of evidence is given explicitly in CE, which makes the evaluation much easier to do, and more reliable.

Using CE in our work has another advantage. As new results of clinical trials are published fairly quickly, we need to provide the latest information to clinicians. We hope that this work will contribute to semi-automatically constructing summaries for CE.

3 Detection of clinical outcomes and their polarity

Clinical outcomes have three general polarities: positive, negative, and neutral. In this subtask, we focus on the problem of detecting the existence of a clinical outcome in medical text, and, when an outcome is found, determining whether it is positive, negative, or neutral, as shown in the following examples.

(1) *Positive*: Patients randomized to receive streptokinase had improved survival compared with those randomized to placebo at 5 years and 12 years.

(2) *Negative*: Meta-analysis of 6 phase 3 trials indicated a significant increase in risk of ICH (intracranial hemorrhage).

(3) *Neutral*: The administration of nifedipine, 30 mg/d, between 7 and 22 days after hospitalization for an acute myocardial infarction (Secondary Prevention Reinfarction Israel Nifedipine Trial study) showed no effect on subsequent mortality and morbidity.

(4) *No outcome*: All patients without specific contraindications were given atenolol (5-10 mg iv) and aspirin (300-325 mg a day).

3.1 Related work

The problem of polarity analysis is also considered as a task of sentiment classification (Pang et al., 2002; Pang and Lee, 2004) or semantic orientation (Turney, 2002). All these tasks are to obtain the orientation of the observed text on a discussion topic. They fall into three categories: detection of the polarity of words, sentences, and documents. Among them, as Yu and Hatzivassiloglou (2003) pointed out, the problem at the sentence level is the hardest one.

Turney (2002) has employed an unsupervised learning method to provide suggestions on documents as *thumbs up* or *thumbs down*. Polarity is determined by averaging the semantic orientation (SO) of extracted phrases from a text. The document is tagged as *thumbs up* if the average of SO is positive, and otherwise tagged as *thumbs down*. Pang et al. (2002) also deal with the task at the document level. Several machine learning techniques are explored to classify movie reviews into positive and negative. They found that unigrams are the most effective lexical feature and are indispensable compared with the other alternatives. The main part of Yu and Hatzivassiloglou's work (2003) is at the sentence level, and hence is closest to our work. They first separate facts from opinions using a Bayesian classifier, then use an unsupervised method to classify opinions as positive, negative, or neutral by evaluating the strength of the orientation of words contained in a sentence.

The polarity information we are observing relates to clinical outcomes instead of the personal opinions studied by the work mentioned above. We expect differences in the expressions and the structures of sentences in these two areas. For the task in the medical domain, we expect that domain knowledge would help. These differences lead to new features as discussed in the following section.

3.2 Method

Support vector machines (SVMs) are used as the classifier to distinguish the four classes in our work. SVMs have been shown to be efficient in text classification tasks (Joachims, 1998). We used the OSU SVM package (Ma et al., 2002) in our experiment.

3.3 Features

In addition to traditional features such as unigrams and bigrams (Pang et al., 2002), we try to improve the performance of classification from two aspects: capture the changes described in the outcome and using generalized features to represent sentences in a more organized way. The following features are explored in our experiment.

3.3.1 Change phrases

Our observation is that outcomes often involve a change in a clinical value (Author1 and Author2, 2004). For example, after a medication was applied to a disease, *mortality* was *increased* or *decreased*. Thus the polarity of an outcome is often determined by how change happens: if a bad thing (e.g., mortality) was reduced, then it is a positive outcome; if the bad thing was increased, then the outcome is negative; if there is no change, then we get a neutral outcome. We try to capture this observation by adding context features.

We manually collected four groups of words: those indicating *more* (*enhanced, higher, exceed, ...*), those indicating *less* (*reduce, decline, fall, ...*), those indicating *good* (*benefit, improvement, advantage, ...*), and those indicating *bad* (*suffer, adverse, hazards, ...*). Two types of features (with the collective name CHANGE PHRASES in the following description) are extracted to address the effects of the changes in different classes. The first emphasizes the effect of words expressing "changes". The way they were added is similar to incorporating the negation effect described by Pang et al. (2002). We attached the tag *_MORE* to all words between the *more*-words and the following punctuation mark, and the tag *_LESS* to the words after the *less*-words. This way, the effect of the "change" words is propagated.

The second class of features addresses the co-occurrence of "change" words and "polarity" words, i.e., it detects whether a sentence expresses the idea of "change of polarity". We use four features for this purpose: MORE GOOD, MORE BAD, LESS GOOD, and LESS BAD. A window of four words on each side of a *more*-word in a sentence is observed to extract the first feature. If a *good*-word occurs in this window, then the feature MORE GOOD is activated. The other three features can be activated in a similar way.

3.3.2 Negations

Negations include expressions with *no* and *not*. We observe that *not* usually does not affect the polarity of a sentence, as shown in the following examples, so they are not included in the feature set.

(5) However, disagreement for uncommon but serious adverse safety outcomes has *not* been examined.

The case for *no* is different: it often suggests a neutral polarity or no clinical outcome at all:

(6) There are *no* short or long term clinical benefits from the administration of nebulised corticosteroids ...

To extract features for negation *no*, all the sentences are first parsed by the Apple Pie parser (Sekine, 1997) to get phrase information for the text. Then, in a sentence containing the word *no*, the noun phrase that *no* is in is extracted. Every word in the noun phrase except *no* itself has a *_NO* tag attached.

3.3.3 Categories

Category information of medical concepts can relieve the data sparseness problem in the learning process. For example, we found that diseases are often mentioned in clinical outcomes as *bad* things:

(7) A combined end point of death or disabling stroke was significantly lower in the accelerated-t-PA group ...

All names of specific diseases in the text are replaced with the tag *DISEASE*.

Intuitively, the occurrences of medical categories, such as intervention and organism function, may be different in the four classes, especially in the *no outcome* class as compared to the other three classes. To verify this intuition, we collected all the categories and use each of them as a feature. Thus, in addition to the words contained in a sentence, all the medical categories mentioned in a sentence are also considered.

The Unified Medical Language System (UMLS) is used as the domain knowledge base, and the software MetaMap (Aronson, 2001) is incorporated for mapping the text to its corresponding medical categories.

3.4 Evaluation

3.4.1 Data set in experiments

We collected 197 abstracts from Medline that were cited in 24 summaries in CE. Every sentence in these abstracts was annotated with the four classes of polarity information. There are 2298 sentences in total.

3.4.2 Results and analysis

The 24 summaries are separated into 10 groups, each containing either 2 or 3 summaries. Ten-fold cross-validation is done by training on 9 groups and testing on the other group. The accuracy is calculated by averaging the 10 results obtained in each run. The features are tested on two tasks. The first task is identification of clinical outcomes (referred to as task1 later). A sentence is classified into two classes: containing a clinical outcome or not. The second task is detection of polarity of outcomes (referred to as task2 later). There are four classes in this task: positive outcome, negative outcome, neutral outcome, or no clinical outcome. An earlier version of task2 was addressed in our previous paper (Author1 et al., 2005 submitted), however, the results reported here are from a different data source and a much larger data set. The results of the two tasks are shown in Table 1.

Not surprisingly, the performance on task1 is better than on task2. For both tasks, the error rates go down as more features are added. The complete feature set has the best performance. With just UNIGRAMS as features, we get 19.97% error rate for task1, which is taken as the baseline. The addition of BIGRAMS in the feature set results in a decrease of 1.3% in the error rate, which corresponds to 6.5% of relative error reduction. Similar improvements are observed in task2. The effectiveness of bigrams in our experiments contradicts the results obtained by Pang et al. (2002) and Yu and Hatzivassiloglou (2003). In their work, adding bigrams does not make much difference, or even is slightly harmful in some cases. Our results agree with the results obtained from another source of medical text (Author1 et al., 2005 submitted). This interesting result indicates that patterns of co-occurrence of words are more regular (i.e., show more commonness in the same class and have less overlaps in different classes) in medical text. The CHANGE PHRASES and CATEGORY features also increase the accuracy. The results validate our intuition that context information and generalizations are important factors in detecting the polarity of clinical outcomes. Also, as there could have been problems caused by overgeneralization, this result provides some evidence of the right degree of the generalization. NEGATIONS only slightly improve the performance in task2. This could be because some of their effect has been captured by bigrams.

Which class is the most difficult to detect, and why? To answer these questions, we further examined the errors in every class. The precision and recall of each class in task2 are shown in Table 2. It is clear in the table that *negative* has the lowest precision and recall. Most errors occur in distinguishing *negative* from *positive* and *no outcome* classes. A sentence is confusing when the change of clinical value is less obvious; for example, the following sentence is incorrectly identified as *positive*.

(8) The mean increase in height in the budesonide group was 11 cm less than in the placebo group (227 vs238 cm, P=0005); ...

In some cases, it turns out that the *no outcome* class is identified as *negative* because of descriptions of diseases.

(9) Lewy body dementia is an insidious impairment of executive functions with Parkinsonism, visual hallucinations, and fluctuating cognitive abilities and increased risk of falls or autonomic failure.

These examples are difficult in that they contain negative expressions (*e.g. increased risk*), yet do not belong to the *negative* class. New features will be needed to identify them correctly.

4 Answer extraction

Various approaches have been tried in multi-document summarization. Lin and Hovy (2002) extended effective single document summarization techniques in an extraction-based multi-document

Table 1: Results of task1 and task2 with different feature sets

RER=Relative Error Reduction (compared to unigrams)

Features	task1		task2	
	Error Rate (%)	RER (%)	Error Rate (%)	RER (%)
(1) UNIGRAMS	19.97	–	24.72	–
(1)+(2) BIGRAMS	18.67	6.50	23.11	6.51
(1)+(2)+(3) CHANGE PHRASES	18.32	8.26	22.45	9.18
(1)+(2)+(3)+(4) NEGATION	18.32	8.26	22.41	9.34
(1)+(2)+(3)+(4)+(5) CATEGORY	17.84	10.67	21.98	11.08

Table 2: Precision and recall of classes in task2

Positive (P/R) (%)	Negative (P/R) (%)	Neutral (P/R)(%)	No Outcome (P/R) (%)
67.67/62.05	45.68/30.33	62.05/53.09	84.02/90.02

summarization system. Schiffman et al. (2002) derived a measure of importance from analyzing a large corpus. It is then combined with some traditional features to rank sentences. In our work, information of clinical outcomes is incorporated in the summarization process.

4.1 Identifying important sentences

Several features that have been shown to be effective in previous document summarization systems are included in our system.

Maximum Marginal Relevancy (MMR): To avoid redundant information being included in the summary. MMR is a measure of "relevant novelty" (Carbonell and Goldstein, 1998). The hypothesis is that information is important if it is both relevant to the topic and least similar to previously selected information — its *marginal relevance* is high. A sentence is represented by a vector of $tf \cdot idf$ values of the terms it contains. The similarity is measured by the cosine distance between two sentences. A parameter λ can be adjusted to give greater or lesser penalty to redundant information. In our experiment, $\lambda = 0.9$ is the best choice. The score of *marginal relevance* is used as a feature in the summarization process (referred to as feature MMR later).

Position: The position of a sentence in an abstract. We calculated it in three ways: (1) in absolute position, sentence i receives the value i^{-1}; (2) the value of sentence i is $i/length$ of the document; (3) a sentence receives value 1 if it is at the beginning (previous to $0.1 \times$ *document length*) of a document, value 3 if it is at the end (after $0.9 \times$ *document length*) of a document, value 2 if it is in between. (2) is the most effective one and it is used in the final experiment.

Sentence length: A score reflecting length of sentences by word counting, normalized by length of longest sentence (Lin, 1999).

Numerical feature: The assumption is that a sentence with numerical values is more informative. Two options were tried: (1) binary value 1 or 0 for whether a sentence contains a numerical value; (2) the number of numerical values in a sentence. We use (1) in the final evaluation as it performs better.

4.2 Approaches

We use the same package of support vector machines as in the previous section to classify whether a sentence should be included in a summary. All the above features are combined with the knowledge of clinical outcome in a classifier. Summaries in our task often contain descriptions of experimental settings in clinical trials and conclusions drawn from them. Therefore, we keep the sequence of sentences the same as in their original documents. Randomly selected sentences are taken as baseline summaries. We evaluate the performance of the summarization using clinical outcomes and their polarity separately.

5 Evaluation

The data set in this experiment is the same as in Section 3.4.1. Again, the ten-fold cross-validation

is performed on the 24 summaries. The average compression ratio of the 24 summaries in CE is 0.25. The system was evaluated using two methods: sentence-level evaluation and ROUGE.

5.1 Sentence-level evaluation

We observed that, generally speaking, the summaries in CE are close to being extracts, and it is usually possible to identify the original Medline abstract sentence upon which each sentence of the CE summary is based. Therefore, we were able to create a benchmark for our system by converting the summaries in CE into their corresponding extracted summary (this is similar to Goldstein et al., 1999). This was done by matching a sentence in the CE summary to the sentence in the abstract which contains most of the key concepts mentioned in that sentence.

5.1.1 Comparison of features

Receiver Operating Characteristics (ROC) analysis, a classic methodology from signal detection theory which has been widely applied in machine learning research (Flach, 2004), is used to evaluate the performance of every single feature in the summarization task. ROC analysis combines metrics *sensitivity* and *specificity* visually using ROC curves.

$$sensitivity = \frac{TP}{TP + FN} = true\ positive\ rate$$

$$specificity = \frac{TN}{TN + FP} = true\ negative\ rate.$$

TP, TN, FP, FN refers to true positives, true negatives, false positives, and false negatives, respectively. *Sensitivity* and *specificity* do not ignore differences between error types; also, they are not dependent on the probability distribution of the data sets. To compare the performance of features, the ROC curves of every feature at different compression ratios are plotted in Figure 1 in the traditional manner. The *x* axis represents the value of $1 - specificity$, and the *y* axis is the value of *sensitivity*. Although compression ratio is not shown explicitly, it becomes larger (less strict) along the curves as *sensitivity* grows.

The diagonal solid line is the purely chance performance. The other four solid lines represent the effects of manually identified clinical outcome, automatically identified clinical outcome, manually identified polarity, and automatically identified polarity.

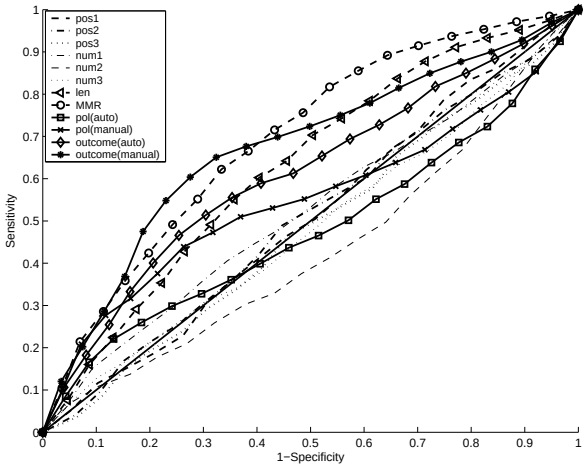

Figure 1: Comparison of features

For a given specificity, the curve which has greater sensitivity will be superior. Similarly, for a given sensitivity, the curve which has greater specificity will be better. This can be observed by checking the area under the ROC curves; a curve is better if it has larger area. It is clear in the figure that knowledge about clinical outcomes helps in this task. At the left part of the figure (compression ratio is smaller), outcome and polarity features are all superior to the purely chance performance. Manually obtained knowledge is even better. It provides good evidence for the importance of using clinical outcomes and polarity. Not surprisingly, MMR is also effective in the task. Other features such as length and position (2) also have good effects on the performance.

5.1.2 Combination of features

When features are combined, some of their effects will be additive, and some will cancel out. The results of different combinations of features at different compression ratios is shown in Table 3. They are obtained by averaging the results of the 24 summaries.

As shown in the table, identification of outcomes and polarity improves the performance at every compression ratio when they are included in the

Table 3: Results of the summarization with different feature sets in sentence-level evaluation

Compression Ratio	0.1			0.2			0.3			0.4		
	P	R	F	P	R	F	P	R	F	P	R	F
Random	.25	.11	.15	.25	.20	.22	.25	.31	.27	.25	.40	.30
MMR	.50	.21	.29	.43	.36	.39	.40	.49	.43	.38	.62	.46
(1) MMR+pos+num+len	.51	.22	.29	.45	.37	.40	.42	.52	.46	.39	.64	.48
(1)+pol (auto)	.48	.21	.29	.47	.39	.42	.42	.53	.46	.41	.67	.50
(1)+pol (manual)	.54	.23	.32	.49	.41	.44	.47	.59	.51	.42	.70	.52
(1)+outcome (auto)	.50	.21	.30	.46	.38	.42	.44	.54	.47	.41	.67	.50
(1)+outcome (manual)	.58	.25	.35	.49	.41	.44	.46	.58	.50	.44	.72	.53

feature set, and there is no big difference between them in the effect. Manually obtained knowledge improves more than automatically obtained knowledge.

The F-score of each summary at compression ratio 0.25 is presented in Figure 2. The diagram at the top shows the results including identification of outcomes, and the one on the bottom shows the results including polarity of outcomes. The figure shows that both of them improves performance in most summaries.

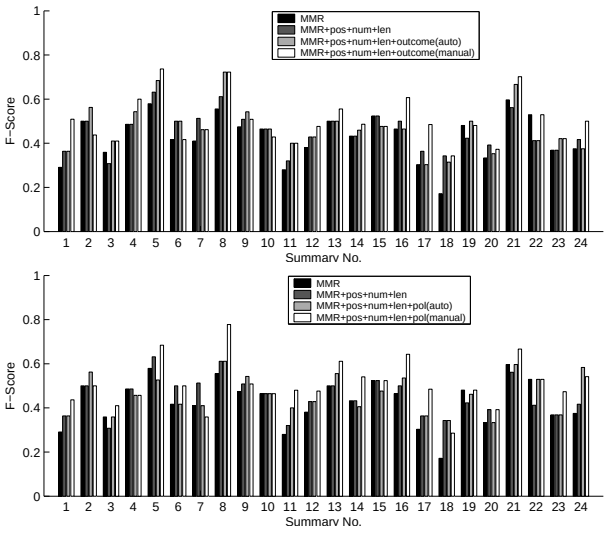

Figure 2: The performance of features in each summary

5.2 ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

The ISI ROUGE package (Lin, 2004) is used in the evaluation. It automatically compares a system-generated summary and a benchmark summary using measures that consider overlapping units such as *n*-grams, word sequences, and word pairs. The results show little difference in the performance of different combination of features. They are briefly listed in Table 4. One reason is that the length feature tends to include longer sentences in the summary in this experiment. When it is combined with outcome or polarity features, more shorter sentences are selected. Although more correct sentences are included in the summary, it is difficult for overlap-based metric to capture this difference.

6 Conclusion and future work

We have described our work of constructing extraction-based summaries to answer questions about effects of applying a medication to a disease. We have shown that domain knowledge and context information is important and effective in identifying clinical outcomes and detecting their polarity. The combined feature set leads to an improvement of about 11% relative error reduction. The ROC curves clearly show that these knowledge is crucial in the summarization task. The evaluation results confirm this conclusion. Comparison of outcome identification and polarity detection shows that there is little difference between them in the effect. However, an additional advantage of using polarity information is that it is necessary in answering polarity-related questions, such as questions asking for side-effects of a medication.

Presentation of the selected sentences in a summary is important in multi-document summarization. Although in our task usually there is no strict

Table 4: ROUGE-L score of different feature sets

Compression Ratio	0.1			0.2			0.3			0.4		
	P	R	F	P	R	F	P	R	F	P	R	F
MMR	.46	.18	.25	.40	.31	.33	.35	.40	.36	.30	.45	.35
MMR+pos+num+len (1)	.46	.18	.25	.41	.31	.34	.35	.40	.36	.30	.45	.35
(1)+outcome (auto)	.46	.18	.25	.41	.31	.34	.35	.39	.36	.30	.46	.35

requirements of ordering sentence (e.g. by time), it is better to reorder sentences from different documents and present them in a coherent way. We would like to investigate how to improve the presentation of the results of summarization in medical domain in the future work.

References

Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In *Proceedings of American Medical Informatics Association Symposium*, pages 17–21.

Author1 and Author2. 2004. Title. In *ACL Workshop Proceedings*.

Author1, Author2, Author3, and Author4. 2003. Title. In *ACL Workshop Proceedings*.

Author1, Author2, Author3, and Author4. 2005, submitted. Title.

Stuart Barton, editor. 2002. *Clinical evidence*. BMJ Publishing Group, London.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Peter A. Flach. 2004. The many faces of ROC analysis in machine learning. In *Tutorial in The Twenty-First International Conference on Machine Learning*.

Thorsten Joachims. 1998. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of the European Conference on Machine Learning (ECML)*, pages 137–142.

Chin-Yew Lin and Eduard Hovy. 2002. From single to multi-document summarization: a prototype system and its evaluation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 457–464.

Chin-Yew Lin. 1999. Training a selection function for extraction. In *Proceedings of the Eighteenth Annual International ACM Conference on Information and Knowledge Management (CIKM)*, pages 55–62.

Chin-Yew Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Workshop on Text Summarization Branches Out*.

Junshui Ma, Yi Zhao, and Stanley Ahalt. 2002. Osu svm classifier matlab toolbox. In *http://www.ece.osu.edu/ maj/osu_svm/*.

Bo Pang and Lillian Lee. 2004. A sentimental education: sentiment analysis using subjectivity smmarizaiton based on minimum cuts. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics*, pages 271–278.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86.

Barry Schiffman, Ani Nenkova, and Kathleen McKeown. 2002. Experiments in multidocument summarization. In *Proceedings of the Human Language Technology Conference*.

Satoshi Sekine. 1997. Apple pie parser. In *http://nlp.cs.nyu.edu/app/*.

Sharon E. Straus and David L. Sackett. 1999. Bringing evidence to the point of care. *Journal of the American Medical Association*, 281:1171–1172.

Peter Turney. 2002. 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*, pages 417–424.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 129–136.