

Using Outcome Polarity in Sentence Extraction for Medical Question-Answering

Yun Niu, MSc, Xiaodan Zhu, MSc, and Graeme Hirst, PhD

Department of Computer Science
University of Toronto
Toronto, Ontario, Canada M5S 3G4

Multiple pieces of text describing various pieces of evidence in clinical trials are often needed in answering a clinical question. We explore a multi-document summarization approach to automatically find this information for questions about effects of using a medication to treat a disease. Sentences in relevant documents are ranked according to various features by a machine-learning approach. Those with higher scores are more important and will be included in the summary. The presence of clinical outcomes and their polarity are incorporated into the approach as features for determining importance of sentences, and the effectiveness of this is investigated, along with that of other textual features. The results show that information on clinical outcomes improves the performance of summarization.

INTRODUCTION

Answers to clinical questions, such as questions posed by clinicians in patient treatment, often require multiple pieces of information. For example:

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

Superficially, this is just a *yes / no* question. But the clinical outcomes of the therapies could be complicated. 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 be obtained only by taking into account various experimental results in the medical research literature and extracting the key points from them. In this paper, we focus on automatically detecting this information for an important type of question: the effects of a medication in the treatment of a disease.

Clearly, identifying clinical outcomes in text is an important component of this task. Moreover, since contradictory evidence can be crucial in answering a clinical question, it will be beneficial to detect not just the presence of an outcome but its *polarity*.

In our earlier work [1], we addressed the problem of detecting outcomes and their polarity in the text of *Clinical Evidence* (see below). In the present paper, we first extend this work to a larger dataset of Medline abstracts, and then investigate its effectiveness in locating potential answers to clinical questions. Given a clinical question and a set of documents that are relevant to it (which may be obtained by information retrieval techniques), we use a multi-document summarization approach to identify information in the documents that is important in answering the question. We observe that a clinical outcome (or, sometimes, two or more outcomes) will usually be described in a single sentence; that is, the description of an outcome does not cross a sentence boundary. Starting with this observation, our goal is to identify a set of sentences as a potential answer to a clinical question.

The prior work most related to ours is on the detection of statements of positive and negative opinion. Yu and Hatzivassiloglou [2] detected sentence-level opinions and their semantic orientation in news articles. Although they mention that polarity information was applied in their system, no details about how it was incorporated were described in their paper, nor was there any evaluation of the effectiveness of this information. Stoyanov et al. [3] analyzed characteristics of opinion questions. Results of some initial experiments showed that filters that identify subjective sentences can be used to guide QA systems involving opinions. A more-detailed review is given in our earlier paper [1].

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 observe that the book *Clinical Evidence* (CE) [4] 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 covers a particular clinical problem, and is divided into subsections that summarize the evidence concerning a particular medication (or a class of medications) for the problem, including results of clinical trials on the benefits and harms. The information sources that CE draws on include medical journal abstracts, review articles, and textbooks. Human experts read the collected information and summarize it to get concise evidence on every specific topic. This is the process of multi-document summarization. Thus, each subsection of CE can be regarded as a human-written multi-document summary of the literature that it cites.

Moreover, we observed that, generally speaking, the summaries in CE are close to being extracts (as opposed to rewritten abstracts). A citation for each piece of evidence is given explicitly, 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. [5]). That is, we matched each sentence in the CE summary to the sentence in the Medline abstract on which it was based (if any) by finding the sentence that contained most of the same key concepts mentioned in the CE sentence.

Using CE in our work has an additional 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-automatic construction of summaries for CE.

DETECTION OF CLINICAL OUTCOMES AND THEIR POLARITY

An earlier version of our work on polarity detection was presented in [1]. In this section, we summarize that work, and report new results from a different data source and a much larger data set.

Clinical outcomes have three general polarities: positive, negative, and neutral. In this subtask, we focus on 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 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).

METHOD

We use support vector machines (SVMs) as the classifier to distinguish the four classes. SVMs have been shown to be efficient in text classification tasks [6]. We used the OSU SVM package [7] in our experiment. The features explored in our experiment are briefly outlined below; details are given in [1].

Unigrams. Words occurring more than 3 times in the data set are extracted as features.

Change phrases. Our observation is that outcomes often involve a change in a clinical value [1]. For example, after a medication is used in the treatment of a disease, mortality might be *increased* or *decreased*. Thus the polarity of an outcome is often determined by how change happens: if a bad thing (e.g., mortality) is reduced, then it is a positive outcome; if the bad thing is increased, then the outcome is negative; if there is no change, then the outcome is neutral. *Change phrases*—phrases that explicitly describe a change in a state or value—are used as features to capture this observation. We manually collected four groups of words (306 in total): 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 change-phrase features are extracted to address the effects of the changes in different classes. In the first type, 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.

- (5) The first systematic review found that β blockers significantly reduced *_LESS* the *_LESS* risk *_LESS* of *_LESS* death *_LESS* and *_LESS* hospital *_LESS* admissions *_LESS*.

The second class of change-phrase 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 recorded. The other three features operate in a similar way.

Bigrams. We also use bigrams (stemmed by Porter’s stemmer [8], occurring more than 3 times in the data set) in the feature set. Although no beneficial effects were observed for bigrams in previous sentiment analysis work, they might help in our task in describing the changes.

Negations. Noun phrases containing the word *no*.

Categories. Semantic types of medical concepts in Unified Medical Language System (UMLS) are used as category features.

EVALUATION OF POLARITY DETECTION

In our new application of this method, we collected 197 abstracts from Medline that were cited in 24 subsections of CE, and annotated each sentence with one of the four classes of polarity information. Each sentence was annotated by one of the first two authors of the paper. There were 2298 sentences in total: 468 expressed positive clinical outcomes, 122 were negative, 194 were neutral, and 1514 did not contain outcomes.

In the experiment, 20% of the data was randomly selected as the testing set and the rest as the training set. The averaged accuracy is obtained from 50 runs. To compare the effectiveness of presence of outcomes and polarity of outcomes in the sentence extraction task, two versions of the task were tried. The first (referred to as task1 below) is simple identification of clinical outcomes: a sentence is classified as either containing a clinical outcome or not. The second task (referred to as task2 below) is detection of polarity of outcomes. There are four classes in this task: positive outcome, negative outcome, neutral outcome, or no clinical outcome.

The results of the two tasks are shown in Table 1. The baseline is randomly assigned labels. Not surprisingly, the performance on task1 is better than that on task2. For both tasks, the error rates go down as more features are added. The complete feature set has the best performance.

Table 1. Evaluation of detection of clinical outcomes and their polarity. Task1 = two-class classification (outcome or not); task2 = four-class classification (positive, negative, neutral, or no outcome).

Approach	task1	task2
	Accuracy (%)	Accuracy (%)
Baseline	65.9	65.9
All Features	82.5	78.3

FACTORS IN IDENTIFYING IMPORTANT SENTENCES

We now turn to our main task, identifying those sentences in a relevant text that would be important to include in an answer to a clinical question. In addition to the presence and polarity of an outcome, as determined by the method described in the previous section, we consider a number of other features that have been shown to be effective in text summarization tasks [9]:

Position of a sentence in an abstract: Sentences near the start or end of a text are more likely to be important. We experimented with three different ways of representing sentence position: option 1, absolute position: sentence i receives the value $i - 1$; option 2, the value for sentence i is $i / \text{length of the document}$; option 3, a sentence receives value 1 if it is at the beginning (first 10%) of a document, value 3 if it is at the end (last 10%) of a document, value 2 if it is in between.

Sentence length: A score reflecting the length of sentences by word counting, normalized by the length of the longest sentence [9].

Numbers: A sentence is more likely to be important if it contains a numerical value. Three options were tried: option 1, binary value 1 or 0 for whether or not the sentence contains a numerical value; option 2, the number of numerical values in the sentence; option 3, binary value 1 or 0 for whether or not the sentence contains the symbol ‘%’.

Maximal Marginal Relevance (MMR): MMR is a measure of “relevant novelty” [10]. Its aim is to find a good balance between relevance and redundancy. 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 sentence vectors. A parameter λ can be adjusted to give greater or lesser penalty to redundant information. The score of marginal relevance is used as a feature in the experiment (referred to as feature MMR).

APPROACHES

We use SVMs, as in the previous section, to rank sentences by their importance values. All the above factors are features in the experiment.

To evaluate the performance of features, the subsections in CE are viewed as ideal summaries of the abstracts that they cite, and the sentences selected as important are compared against them for evaluation,

as described below. Hereafter, we use *summaries* to refer to the two sets.

EVALUATION

The data set in this experiment is the same as in the polarity detection task: 197 abstracts cited in 24 subsections (summaries) in CE are used. The average compression ratio (*number of sentences in a summary divided by total number of sentences in the original abstracts that are cited in the summary*) of the 24 summaries in CE is 0.25. Out of the total 2298 abstract sentences, 784 contain a clinical outcome (34.1%). The total number of sentences in the 24 summaries is 546, of which 295 sentences contain a clinical outcome (54.0%).

Average results are calculated over 50 runs of randomly selecting 3 summaries as the testing data and the other 21 as the training set. As the purpose is to observe the behavior of different feature sets, the experimental process can be viewed as a glass box. The system was evaluated by two methods: sentence-level evaluation and ROUGE. Randomly selected sentences are taken as baseline summaries.

Sentence-Level Evaluation

Comparison of individual features

The precision and recall curves of every feature at different compression ratios are plotted in Figure 1. The purely chance performance has precision of 0.25, shown by the solid horizontal line. The other four solid lines represent the effects of manually or automatically identified clinical outcome and polarity. Although compression ratio is not shown explicitly in the figure, lower compression ratios correspond to lower recall (left-hand part of the figure). It is clear in the figure that knowledge about clinical outcomes helps in this task. On the left part of the figure, outcome and polarity features are all superior to the baseline performance. Manually obtained knowledge is even better. Not surprisingly, MMR is also effective in the task. Other features such as length and numerical value (option 1) also have good effects on the performance.

Combining the features

When features are combined, some of their effects will be additive, and some will cancel out. Table 2 shows the results, in terms of precision, recall, and F -score, from different combinations of features at different compression ratios. (To save space, only the best results of MMR ($\lambda = 0.9$), position (option 2), and numerical features (option 1) are listed.)

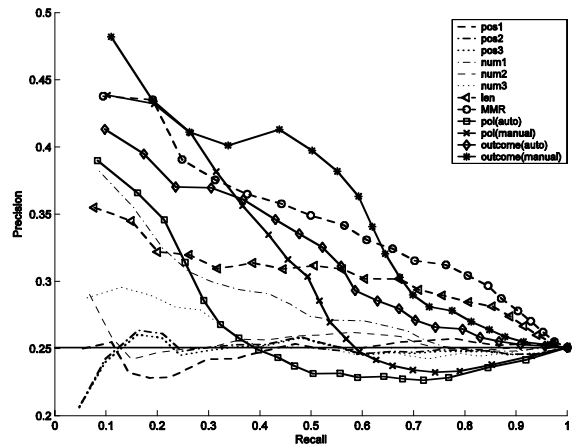


Figure 1. Precision and recall curves for individual features on the task of selecting sentences for a summary.

The results show that including the identification of outcomes and polarity in the feature set improves the performance by up to 5 points in F -score at every compression ratio. But the additional benefit from further determining the polarity of the outcome rather than just detecting the presence or absence of outcomes is small. We had expected that, intuitively, polarity may provide more information on contradiction and thus may help more in the task. Observing the data, however, we found that one aspect accounting for the result could be that although some sentences are different in polarity, they do not form contradictions. Rather, for example, they describe different clinical outcomes, and some of the outcomes are not important and thus are not included in the summaries. Again, manually obtained knowledge gives a larger improvement than automatically obtained knowledge.

ROUGE

As an alternative evaluation, we use the ISI ROUGE package [11], which compares a summary generated by a text summarization system with a benchmark summary by considering overlapping units such as n -grams, word sequences, and word pairs. Our evaluation was carried out with various ROUGE parameters. Unlike the sentence-level evaluation, the results showed little difference in the performance of different combination of features. One reason could be that it is difficult for an overlap-based metric to capture the difference if the content of two sets is similar. For example, only a small difference might be measured by ROUGE when comparing the inclusion of both a positive and a negative clinical outcome (of the same medication in treatment of the same disease) in the summary with the inclusion of only one of them.

Table 2: Sentence-level evaluation of the summarization at different compression ratios with different feature sets. Best results for each compression ratio are shown in boldface. P = precision; R = recall; F = F-score.

Compression Ratio	0.1			0.2			0.3			0.4		
	P	R	F	P	R	F	P	R	F	P	R	F
Baseline	.25	.11	.15	.25	.20	.22	.25	.31	.27	.25	.40	.30
MMR	.44	.19	.27	.38	.31	.34	.36	.44	.39	.34	.57	.42
(1) MMR+pos+num+len	.44	.19	.26	.40	.33	.36	.38	.48	.42	.36	.58	.44
(1)+polarity (auto)	.45	.20	.27	.42	.35	.38	.39	.49	.43	.37	.61	.46
(1)+polarity (manual)	.49	.21	.29	.44	.38	.40	.41	.52	.46	.38	.64	.48
(1)+outcome (auto)	.45	.20	.27	.41	.35	.38	.39	.48	.43	.37	.61	.46
(1)+outcome (manual)	.51	.22	.31	.45	.38	.41	.42	.53	.46	.39	.65	.48

CONCLUSION

We have described our work on identifying important sentences to answer questions about effects of using a medication to treat a disease. We have shown that combining context information and domain knowledge achieves the best performance in identifying clinical outcomes and detecting their polarity. The accuracy in the two tasks achieved by combining all features is 82.5% and 78.3% respectively. This information is incorporated into a multi-document summarization approach to locate potential answers. Results of comparison and combination of features clearly show that detecting the presence and polarity of clinical outcomes is helpful in importance evaluation. However, using automatically detected polarity information results in only a slight improvement in the results of the importance evaluation task. Thus, our next step will be to build more accurate polarity detection systems. An additional advantage of having polarity information is that some questions require it in their answer, such as questions asking for harmful side-effects of a medication.

REFERENCES

- [1] Niu Y, Zhu XD, Li JH, Hirst G. Analysis of polarity information in medical text. In: Proceedings of the American Medical Informatics Association 2005 Annual Symposium; 2005. p. 570–574.
- [2] Yu H, Hatzivassiloglou V. 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; 2003. p. 129–136.
- [3] Stoyanov V, Cardie C, Wiebe J. Multi-perspective questions answering using the OpQA corpus. In: Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing; 2005. p. 923–930.
- [4] Barton S, editor. Clinical evidence. London: BMJ Publishing Group; 2002.
- [5] Goldstein J, Kantrowitz M, Mittal V, Carbonell J. Summarizing text documents: Sentence selection and evaluation metrics. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and Development in Information Retrieval; 1999. p. 121–128.
- [6] Joachims T. Text categorization with support vector machines: learning with many relevant features. In: Proceedings of the European Conf on Machine Learning; 1998. p. 137–142.
- [7] Ma J, Zhao Y, Ahalt S. OSU SVM Classifier Matlab Toolbox; 2002.
- [8] Porter MF. An algorithm for suffix stripping. Program 1980; 14(3):130–137
- [9] Lin CY. Training a selection function for extraction. In: Proceedings of the 18th Annual International ACM Conference on Information and Knowledge Management; 1999. p. 55–62.
- [10] Carbonell J, Goldstein J. 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; 1998. p. 335–336.
- [11] Lin CY. ROUGE: a package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out, 42nd Annual Meeting of the Association for Computational Linguistics; 2004. p. 74–81.