

Analysis of Polarity Information in Medical Text

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

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

Knowing the polarity of clinical outcomes is important in answering questions posed by clinicians in patient treatment. We treat analysis of this information as a classification problem. Natural language processing and machine learning techniques are applied to detect four possibilities in medical text: no outcome, positive outcome, negative outcome, and neutral outcome. A supervised learning method is used to perform the classification at the sentence level. Five feature sets are constructed: UNIGRAMS, BIGRAMS, CHANGE PHRASES, NEGATIONS, and CATEGORIES. The performance of different combinations of feature sets is compared. The results show that generalization using the category information in the domain knowledge base Unified Medical Language System is effective in the task. The effect of context information is significant. Combining linguistic features and domain knowledge leads to the highest accuracy.

INTRODUCTION

A crucial issue in searching text to answer clinical questions is identifying the polarity of clinical outcomes: was the outcome “good” or “bad”? We focus on the problem of detecting the presence 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) *No outcome:* We found no RCTs comparing combined pharmacotherapy and psychotherapy with either treatment alone.
- (2) *Positive:* Thrombolysis reduced the risk of death or dependency at the end of the studies.
- (3) *Negative:* In the systematic review, thrombolysis increased fatal intracranial haemorrhage compared with placebo.
- (4) *Neutral:* The first RCT found that diclofenac plus misoprostol versus placebo for 25 weeks produced no

significant difference in cognitive function or global status.

The context for our work is the EPoCare project (“Evidence at Point of Care”) at the University of Toronto, which is developing point-of-care online access to clinical evidence. Clinicians often need to consult literature on the latest information in patient care, such as side effects of a medication, symptoms of a disease, or time constraints in the use of a medication [1, 2]. For example:¹

- Q:** In a patient with a suspected MI does thrombolysis decrease the risk of death if it is administered 10 hours after the onset of chest pain?
- Q:** In a patient with a generalized anxiety disorder, does cognitive behavior or relaxation therapy decrease symptoms?

While the answers to these questions could be found by analyzing and consolidating experimental results in the research literature, an alternative source is *Clinical Evidence* (CE) [3], a regularly updated publication that reviews and consolidates experimental results for clinical problems. The following text in CE is relevant to the questions:

- A:** Systematic reviews of RCTs have found that prompt thrombolytic treatment (within 6 hours and perhaps up to 12 hours and longer after the onset of symptoms) reduces mortality in people with AMI and ST elevation or bundle branch block on their presenting ECG.
- A:** Two systematic reviews have found that cognitive therapy, using a combination of behavioural interventions such as exposure, relaxation, and cognitive restructuring, improves anxiety and depression more than remaining

¹ These examples are 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 Toronto.

on a waiting list (no treatment), anxiety management training alone, or nondirective treatment.

The present work forms part of our research to develop methods for automatically answering questions with CE as the source text. Polarity information is a necessary part of this for several reasons. First, we need to know the polarity to answer questions about benefits and harms of an intervention. Second, the case of *no outcome* helps filter out irrelevant information when the question is asking about the clinical outcomes of an intervention. Third, negative outcomes describing side effects may be crucial for a clinical decision even if the question does not require it explicitly. Finally, from the number of positive or negative descriptions of the outcome of an intervention applying to a disease, clinicians can form a general idea about how “good” the intervention is.

We describe our work of applying natural language processing and machine learning techniques for this task. The results show that combining linguistic features and domain features achieves the best performance, with accuracy of 79.42%.

BACKGROUND

The problem of polarity analysis is also considered as a task of sentiment classification [4, 5] or semantic orientation [6]: determining whether an evaluative text, such as a movie review, expresses a “favorable” or “unfavorable” opinion. 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 [7] pointed out, the problem at the sentence level is the hardest one.

Turney [6] has employed a 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*. The SO of a phrase is calculated as the difference in the mutual information between an observed phrase and the positive word *excellent*, and the observed phrase and the negative word *poor*. Documents are classified as either positive or negative; no neutral position is allowed.

Pang et al. [4] also deal with the task at the document level. Several machine learning techniques are explored to classify movie reviews into positive and negative. Three classification strategies, naive Bayes, maximum entropy classification, and support vector machines are investigated. Meanwhile, a series of

lexical features including unigrams, bigrams, and part-of-speech tags are employed in these classification strategies in order to find effective features. Pang et al. found that machine learning techniques can always outperform a human-generated baseline; among the three classification strategies, support vector machines perform the best; unigrams are the most effective lexical feature and indispensable compared with the other alternatives.

The main part of Yu and Hatzivassiloglou’s work [7] 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, and 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, it will be interesting to see if domain knowledge would help. These differences lead to new features as discussed in the following section. We define our task as a four-way classification problem: no outcome, positive outcome, negative outcome, and neutral outcome. We apply a supervised method to classify the four classes in a uniform way.

METHOD

A support vector machine (SVM) is used as a classifier to distinguish the four classes in our work. SVMs have been shown to be efficient in text classification tasks of natural language processing [8]. Given a training set, the SVM finds a hyperplane with the maximal margin of separation between two classes. The classification is then just to determine which side of the hyperplane the test sample lies in. We used the OSU SVM package [9] in our experiment.

Features

In order to do the classification, we need to extract features that reflect the difference between the classes. Some features are selected according to the linguistic characteristics in the expression of the text.

Unigrams and Bigrams

As different words are expected to appear in sentences of the four classes, we add two types of word features: UNIGRAMS and BIGRAMS to the feature set. To obtain unigram features, we extract every single word that occurs more than three times in the training text. For bigram features, every two adjacent words in the text are combined in order to catch some

word patterns appearing commonly in a class. As for the UNIGRAMS, only BIGRAMS with frequency more than three are extracted. We use stemmed words (obtained by Porter's stemmer [10]) when extracting BIGRAMS.

Change Phrases

Our observation is that outcomes often involve a change in a clinical value. For example, after a medication was applied to a disease, *mortality was increased or decreased* [11].

- (5) In these three postinfarction trials ACE inhibitor versus placebo significantly reduced mortality, readmission for heart failure, and reinfarction.

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 same 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. [4]. 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.

- (6) 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*.
- (7) Another large rct found milrinone versus placebo increased *_MORE mortality_MORE over_MORE 6_MORE months_MORE*.

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. These features are designed mostly to distinguish between positive, negative, and neutral cases.

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 we do not take it into account in the feature set.

- (8) The first RCT found fewer episodes of infection while taking antibiotics than while not taking antibiotics.
- (9) The rates of adverse effects seemed higher with rivastigmine than with other anticholinesterase drugs, but direct comparisons have not been performed.

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

- (10) One systematic review in people with Alzheimer's disease found no significant benefit with lecithin versus placebo.
- (11) We found no systematic review or RCTs of rivastigmine in people with vascular dementia.

To extract features for negation *no*, all the sentences are first parsed by the Apple Pie parser [12] 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.

Categories

Other features are based on the category information of medical concepts in a medical knowledge base.

Category information can relieve the data sparseness problem in the learning process. All names of specific diseases in the text are generated to the *disease* category by replacing them with the tag *DISEASE*.

Intuitively, the occurrences of semantic types, such as *disease or syndrome* 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 types that occur in the training text and use each of them as a feature. Thus, in addition to the words contained in a sentence, all the semantic types mentioned in a sentence are also considered.

The Unified Medical Language System (UMLS) is used as the domain knowledge base, and the software MetaMap [13] is incorporated for mapping the text to their corresponding concepts in the UMLS

Metathesaurus. The semantic type of a concept is then extracted.

Data Set

The data set with four classes was built by collecting sentences from different sections in CE; 1509 sentences were used (472 positive, 338 negative, 250 neutral, 449 none). All examples were labeled manually.

TRAINING

We randomly select 20% data of the whole data set as the test set (301 sentences), and use the rest (1208 sentences) as the training set.

In the training process, we gradually add training samples until all of them are included, and observe the performance on test set.

RESULTS

The results are shown in Figure 1. As the figure indicates, the error rates go down as more training data is used, and when more features are added. The complete feature set performs consistently the best. The results match our intuition that context information and generalizations are important factors in detecting the polarity of clinical outcomes.

The results of the five feature sets applied in the classification (using the full training set) are shown in Table 1. With just UNIGRAMS as features, we get 25.12% error rate, which is taken as the baseline. The addition of BIGRAMS in the feature set results in a decrease of about 3% in the error rate, which corresponds to 11.9% of relative error reduction.

DISCUSSION

The effectiveness of bigrams in our experiments contradicts the results obtained by Pang et al. [4] and Yu and Hatzivassiloglou [7]. In their work, adding bigrams does not make much difference, or even is slightly harmful in some cases. 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 the medical text we are observing. The CATEGORY features also increase the accuracy. The relative error reduction obtained by adding this set is about 3% percent. This shows that generalization is important in this task. Also, as there could have been problems caused by over-generalization, this result provides some evidence of the right degree of the

generalization. CHANGE PHRASES and NEGATIONS only slightly improve the performance. This could be because that some of their effect has already been captured by bigrams.

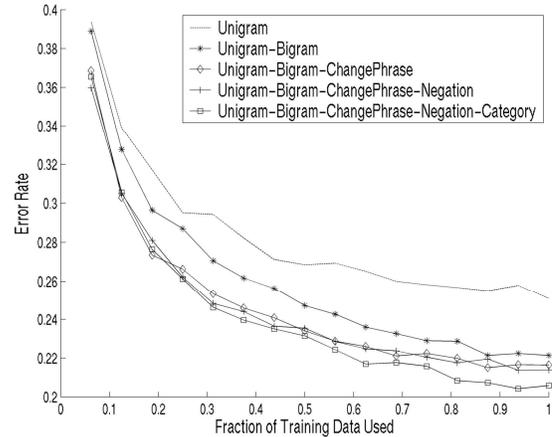


Figure 1. Error rate of classes using different fractions of training data

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 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. In the former case, a sentence is confusing when it has phrasings that seem to contrast; for example, the following sentence is incorrectly identified as *negative*.

- (14) Despite the frequent adverse effects, people receiving active treatment were more likely to stay in trials than those receiving placebo in both the short and the medium term.

As for the latter case, it turns out that descriptions of diseases in the *no outcome* class are often identified as *negative*.

- (15) 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 (*adverse effects, increased risk...*), yet do not belong to the *negative* class. New features will be needed to identify them correctly.

Table 1. Results of the classification with different feature sets

Features	Error Rate (%)	Relative Error Reduction (%) (over Unigrams)
(1) UNIGRAMS	25.12	—
(1)+ (2) BIGRAMS	22.13	11.9
(1)+ (2)+ (3) CHANGE PHRASES	21.64	13.9
(1)+ (2)+ (3)+ (4) NEGATION	21.38	14.9
(1)+ (2)+ (3)+ (4)+ (5) CATEGORY	20.58	18.1

Table 2. Precision and recall of classes

Classes	Positive	Negative	Neutral	No Outcome
Precision (%)	86.81	73.13	79.16	76.84
Recall (%)	83.15	73.13	76.00	82.02

CONCLUSION

We have described our work in the analysis of the polarity of clinical outcomes in medical text. We have shown that the combination of linguistic features and domain knowledge features leads to good performance in classifying the four cases: no outcomes, positive outcomes, negative outcomes, and neutral outcomes. Our next step is to work on summarizing polarity information in published literature based on the results of classification.

REFERENCES

- [1] Sackett DL, Straus SE. Finding and applying evidence during clinical rounds: the “evidence cart”. *Journal of the American Medical Association* 1998;280(15):1336–1338.
- [2] Straus SE, Sackett DL. Bringing evidence to the point of care. *Journal of the American Medical Association* 1999;281:1171–1172.
- [3] Barton S, editor. *Clinical evidence*. London: BMJ Publishing Group; 2002.
- [4] Pang B, Lee L, Vaithyanathan S. Thumbs up? sentiment classification using machine learning techniques. In: *Proceedings of 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; 2002. p. 79–86.
- [5] Pang B, Lee L. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics*; 2004. p. 271–278.
- [6] Turney P. 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*; 2002. p. 417–424.
- [7] 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.
- [8] Joachims T. Text categorization with support vector machines: learning with many relevant features. In: *Proceedings of the European Conference on Machine Learning (ECML)*; 1998. p. 137–142.
- [9] Ma J., Zhao Y., Ahalt S. OSU SVM Classifier Matlab Toolbox; 2002. Available from: URL: http://www.ece.osu.edu/~maj/osu_svm/.
- [10] Porter MF. An algorithm for suffix stripping. *Program* 1980;14(3):130–137.
- [11] Niu Y, Hirst G. Analysis of semantic classes in medical text for question answering. In: *Proceedings of 42st Annual Meeting of the Association for Computational Linguistics, Workshop on Question Answering in Restricted Domains*; 2004. p. 54–61.
- [12] Sekine S. Apple Pie Parser; 1997. Available from: URL:<http://nlp.cs.nyu.edu/app/>.
- [13] Aronson AR. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In: *Proceedings of American Medical Informatics Association Symposium*; 2001. p. 17–21.