Detection of Evidence in Clinical Research Papers

Patrick Davis-Desmond **Diego Mollá**

Department of Computing
Macquarie University
Sydney 2109, NSW, Australia
Email: diego.molla-aliod@mq.edu.au

Abstract

When appraising published clinical research, medical doctors and researchers often need to know whether the clinical outcomes presented had statistical evidence. In this paper we present a study for the detection of expressions of such statistical evidence. An effective rule-based classifier has been developed that uses regular expressions and a list of negation phrases to automatically classify documents as either showing evidence of effect in the results or not. The classifier performed with an accuracy between 88% and 98% at 95% confidence intervals, and it also outperformed a set of baselines using bag-of-word features in several statistical classifiers. The rule-based system is written in Python and is available as open-source code.

Keywords: evidence-based medicine, appraisal, text classification.

1 Introduction

On-line medical databases, such as PubMed[1] and PubMed Central[2] maintained by the US National Library of Medicine, publish thousands of clinical papers yearly. These free full-text reports can be retrieved to find relevant information for clinical and research purposes. Searching through this research literature to find relevant articles and the best evidence for questions posed by medical practitioners can be a daunting and time-consuming task. Many of the articles retrieved will be irrelevant in cases where the research hypothesis has been rejected due to lack of clinical evidence and statistical proof. The reports most needed by medical practitioners to help find answers to clinical questions are the ones where clinical evidence has been found and the research hypothesis has been accepted.

Current information retrieval systems used to research these on-line databases fail to differentiate between articles that have clinical evidence or not. As a result, many articles retrieved are not relevant to the medical practitioner. Detection of lack of evidence will allow the bulk of non-relevant articles to be excluded from the search results. Medical practitioners can then focus on reading articles with proven clinical evidence in order to find information of benefit to their patients.

In this paper we present an initial study on the detection of evidence in published medical research papers. We focus on Randomised Controlled Trials (RCT) and show that a rule-based approach that targets the detection of specific expressions of negation in the text gives an accuracy between 88% and 98% at 95% confidence intervals.

The structure of the following sections of this paper is as follows. Section 2 presents work on aspects related to the detection of clinical evidence. Section 3 details the methodology that we have followed. In particular, Section 3.1 shows how the corpus of RCTs has been gathered, Section 3.2 focuses on how the corpus was annotated, Section 3.3 presents a set of baselines that we developed using machine learning methods on bag-of-word features, and Section 3.4 details our rule-based system. Finally, Section 4 presents the conclusions.

2 Related Work

We are not aware of any work that specifically targets the detection of clinical evidence, but there has been work on the detection of polarity of clinical outcomes and substantial work on various tasks related to the detection of negation in clinical texts.

Niu et al. (2006) showed an improvement of the results of a multi-document summarisation approach over clinical trials by incorporating information on the polarity of clinical outcomes. Their polarity detection system classified the clinical studies into one of four types according to whether the outcomes improved the patient outcomes: "positive", "negative", "neutral", "no outcome".[3] They applied SVM on a corpus of 197 abstracts and obtained a maximum accuracy of 82.5%.

Preliminary visual inspection of our corpus of research papers revealed the occurrence of a substantial number of negation expressions that indicate lack of evidence in their findings. We therefore focused our work on detecting those expressions automatically. There is a number of approaches researching negation detection applied to medical texts. One of the better known studies is NegEx (Chapman et al., 2001). NegEx is based on regular expressions and its

The work presented here is based on an extension of coursework of the first author at the Masters of Information Technology, Macquarie University.

[1] http://www.ncbi.nlm.nih.gov/pubmed/

[2] http://www.ncbi.nlm.nih.gov/pmc/

[3] The difference between "neutral" and "no outcome" is whether the results indicated no significant evidence ("neutral") or whether the study did not provide the results ("no outcome").

focus is on detecting negated findings and diseases in discharge summaries. The algorithm uses several phrases indicating negation and filters out sentences containing "pseudo-negations", that is phrases that falsely appear to be negation phrases such as double negatives and ambiguous phrasing (e.g. *unremarkable*). The system also limits the scope of negation phrases to a context window with size of five words either side of the concept. Their algorithm uses a predefined set of pseudo-negation phrases, a set of negation phrases, and two simple regular expressions. NegEx performed with an accuracy of about 84% and a recall of about 78%. The rule-based negation classifier developed for our project is a modified and simplified version of NegEx.

There have been several other development efforts based on NegEx. Skeppstedt (2010) evaluated NegEx on clinical health records in Swedish and achieved a precision of 70% and a recall of 81% for sentences containing the negation phrases. Although the results are not as high as results achieved by NegEx, Skeppstedt comments that "the comparison between the English and Swedish evaluations is complicated by the fact that the Swedish test data had lower inter-rater agreement", which is likely to have affected the Swedish results. The author also notes certain language specific examples such as *icke* (meaning *not* or *non-*), which commonly appears at the start of the disease names like *icke allergisk astma*. NegEx interprets this as negation, thus affecting precision. Goryachev et al. (2006) evaluated four different methods of negation detection, including regular-expression-based algorithms, syntactic-processing-based algorithms, and statistical classifiers including Naïve Bayes and SVM. They modified NegEx and another negation algorithm called NegExpander, the latter developed by Aronow et al. (1999). The study reported a result of 92% accuracy with their modified version of NegEx, which they applied to hospital outpatient reports. Their conclusion was that rule-based classifiers performed better than statistical classifiers. Meystre and Haug (2005) created a modified version of NegEx, so as to detect negation when extracting medical problems from medical records. The negation detection algorithm of NegEx was also modified to match UMLS concepts contained in a keyword table. The aim of Meystre's work was to develop a natural language processing tool that would harvest potential "problem list" entries from the electronic documents in the LDS Hospital (Salt Lake City, UT) Electronic Medical Record system. These documents comprised free-text documents of patient medical history and reports of medical interventions or clinical progress. The best results were obtained using their MMTx2 tool (75.3% precision and 89.2% recall).

There has been additional work on the detection of negation besides NegEx and its variants. NegExpander (Aronow et al., 1999) was designed to detect expressions of observed evidence[4] in radiology reports and uses syntactic processing techniques to identify noun phrases or conjunctive phrases that define negation boundaries. NegFinder (Mutalik et al., 2001) uses regular expressions and a parser to identify negation in hospital discharge summaries and surgical notes. Mutalik et al. noted that "MEDLINE indexing uses sophisticated syntactic and semantic processing techniques, but does not incorporate explicit distinctions between positive and negative terms". The algorithm finds negated concepts in discharge summaries and surgical notes with 91.8% accuracy and 95.7% recall. ChartIndex (Huang and Lowe, 2007) uses regular expressions and parse trees to locate negated medical concepts in radiology reports. ChartIndex was developed in order to automatically extract meta-data as part of the STRIDE (Stanford Translational Research Integrated Database Environment) system from the over one million full-text pathology reports stored in the STRIDE Clinical Data Warehouse (CDW). ChartIndex achieved accuracy results between 85% and 92%.

Recent work by Uzuner et al. (2009) compared machine learning and rule-based approaches to classifying discharge summaries, and achieved better results using a statistical classifier (SVM) compared to their own extended version of NegEx. Contextual features, including simple lexical information and more complex syntactic information, are extracted from the text and then used by the statistical classifier. A limited word window of 4 words either side of the target is used to limit the scope of the classifier. For example, the verb *showed* preceding a problem suggests that the condition is present, whereas *cured* after a problem suggests that the condition is absent. Uzuner et al. showed that their statistical classifier, StAC, can capture what they termed "assertion classes" on discharge summaries and radiology reports by making use of the information contained in the immediate context of target problems. Uzuner achieved F-value results of 98% for the positive class and 95% for the negative class using StAC, and 93% for the positive class and 90% for the negative class when using their extended version of NegEx.

Our work differs from those of related work in that we aim specifically at detecting the existence of *statistical evidence* by means of detecting specific negation phrases. Our difference from Niu et al. (2006)'s work is subtler. Whereas Niu et al. Focused on detecting outcomes that were beneficial ("positive") or not ("negative") for the patient, we focus on whether there are any outcomes at all, and, in subsequent work, we will look at detecting the direction of the outcomes. We believe that providing a first classification between outcome/no outcome before detecting its polarity has the potential of giving better results, since the two classifiers can focus on different types of information.

3 Method

3.1 Corpus Gathering

All medical research articles used in this research were sourced from PubMed Central. We selected a specific type of clinical study named Randomised Control Trial (RCT) since they are frequently used to report the results of clinical studies. RCTs are high-quality studies focusing on the generation of measurable outcomes. In a RCT the subjects are randomly allocated to one of two groups. In the "active" group, the subjects are given the treatment that is the object of study; in the "control" group, they are given a placebo. The outcome of a RCT would normally indicate whether there is a statistically significant difference between the results of the active group against the control group.

The research articles were gathered from PubMed Central and stored into a database. Since PubMed Central does not identify RCTs, we did a first pass using PubMed.

The entire procedure was as follows.[5]

[4] Observed evidence is not to be confused with statistical evidence. Our work focuses on the detection of statistical evidence.

[5] Note that the data were gathered around September 2010. The interfaces to PubMed and PubMedCentral may have changed since

1. Go to PubMed and visit the "Limits" section.[6]
2. Select "Published in the last 180 days" in the Dates list, select "Randomized Controlled Trial" as the type of article, select "English" as the language, and select "Links to Free Full-text" as the text option. Click "Search".
3. When the results appear, click the link to "Free Full Text".
4. Visually inspect the list of results to find those that are marked as completed RCTs.
5. Copy the PMID to the database and then click the "Free Text" link to open the article.
6. Identify the PMCID and the PICO details and copy them to the database (see below).
7. Save the record to the database.
8. Go to PubMed Central.[7]
9. Enter the PMCID in the search box, click the "Search all Articles" option, and then click "Search".
10. Click the "Display" link and then select XML to change to display to XML format.
11. Click the "Send To" link and then select "File" to save the file.
12. Save the file using the PMCID as the filename. Also change the extension from .txt to .xml.

Step 6 involved the extraction of the PICO details. These include four key aspects of patient care (Gosall and Gosall, 2009):

1. **P**roblem or patient;
2. The main **I**ntervention, exposure, test or prognostic factor under consideration;
3. A **C**omparative intervention used in treatment; and
4. The **O**utcomes achieved or measured.

The PICO information was extracted by visual inspection of the text and was stored in the database. The purpose of keeping this information was to help the annotators understand the abstracts quicker.

3.2 Corpus Annotation

The abstracts of the RCTs were visually inspected by annotators to determine the type of evidence presented. Three annotators from two medical institutions[8] were recruited as domain experts and the annotation was performed using a web-based annotation tool designed for this annotation task.

The abstract and the PICO details were uploaded to the annotation tool. The annotation tool was designed to allow an arbitrary number of domain experts to annotate the texts.

The annotators were instructed to examine the statistics reported in the abstracts in order to decide whether the research hypothesis had been rejected or not. In particular, if no difference is found between the results of the active and the control group in a RCT, then there is no evidence to support that the intervention under study has any effect in the measured outcomes, and therefore the research hypothesis is deemed as rejected. If, however, statistical differences are found that are not due to chance alone, then the research hypothesis is accepted.

The annotators were therefore required to read the abstract, and then to select either "Accepted", "Rejected", or "Unknown" according to these criteria:

Accepted: A difference is reported between the intervention and the control group.

Rejected: No difference is reported.

Unknown: Unable to tell (e.g. because the RCT does not provide any results).

Figure 1 shows a screen-shot of the summary page presented to the annotators. The page contains the PICO information together with the result of the annotation. The annotators had access to the full abstract as shown in Figure 2.

Abstracts marked as "Unknown" were discarded for this study. The dataset included additional annotations, including annotations about whether there were secondary outcomes and their types. This information will be used in future studies but was discarded in the present study.

Whenever there was disagreement between annotators we asked them to review the articles. The annotators were not influenced to select any class or to change their original classification.

To measure the final agreement between annotators we computed Fleiss' Kappa (κ). The Kappa statistic can be interpreted as expressing the extent to which the observed amount of agreement among annotators exceeds what would be expected if all annotators made their ratings completely randomly. If agreement is no more than expected by chance, then $\kappa = 0$. With perfect agreement, κ=1. The exact formula is

$$\kappa = (P_O - P_E)/(1 - P_E)$$

where P_O is the observed agreement, and P_E is the agreement expected by chance.

We found a Kappa value of 70.6%. This falls within the range of values that is usually termed as "good agreement beyond chance". For the final dataset, we chose the decisions that corresponded to the majority of the annotators' individual decisions.

Table 1 shows the numbers of articles used for training and testing for each type. We can observe that the ratio between evidence ("Accepted") and no evidence ("Rejected") is roughly equal, which approximates the ratio observed in our pilot studies.

	Accepted	Rejected	Total
(1) Training	66	61	127
(2) Test	33	34	67
(1)+(2) Total	99	95	194

Table 1: Dataset used for training and testing

3.3 Baselines

We ran several statistical classifiers using word-based features. We partitioned the corpus into a training set as shown in Table 1. The size of the corpus, though similar to that of related work such as by Niu et al. (2006), was small and we would expect better results

then.

[6] http://www.ncbi.nlm.nih.gov/pubmed/limits

[7] http://www.ncbi.nlm.nih.gov/pmc/

[8] Kolling Medical Research Institute and Royal North Shore Hospital, Sydney.

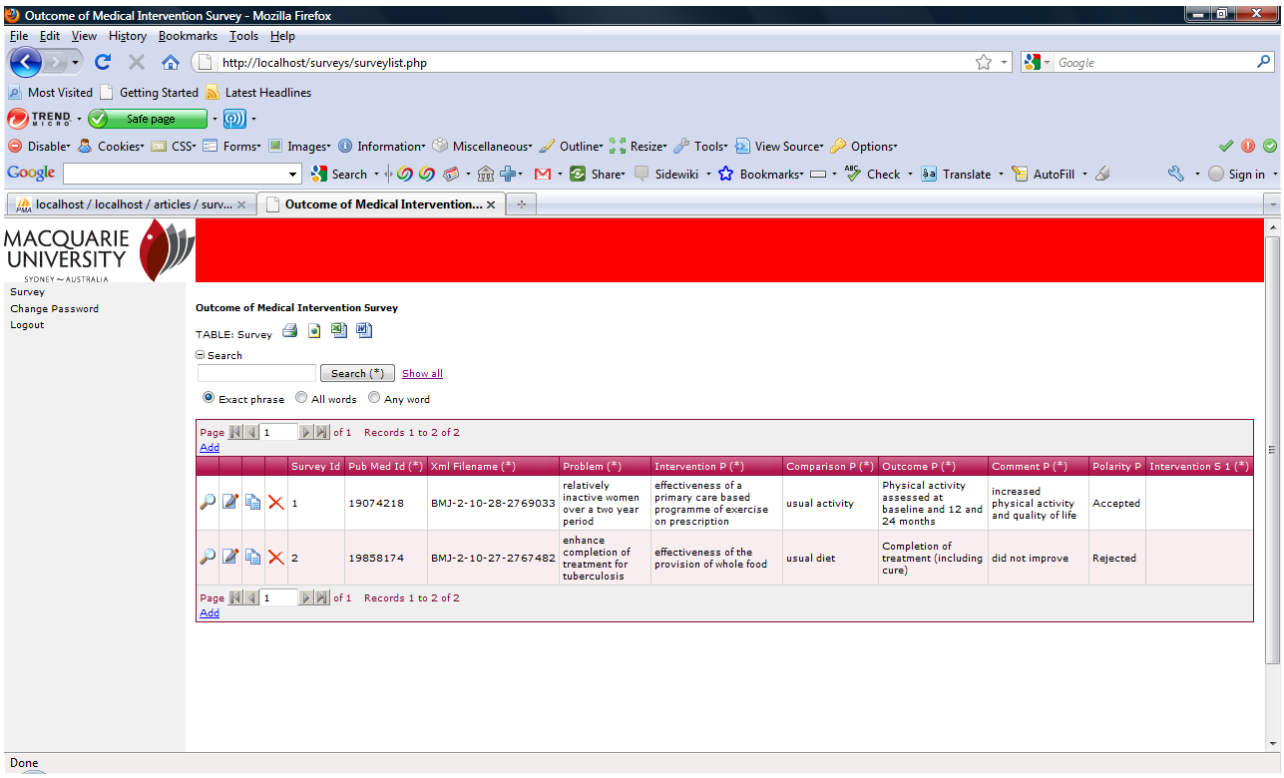

Figure 1: Summary listing page

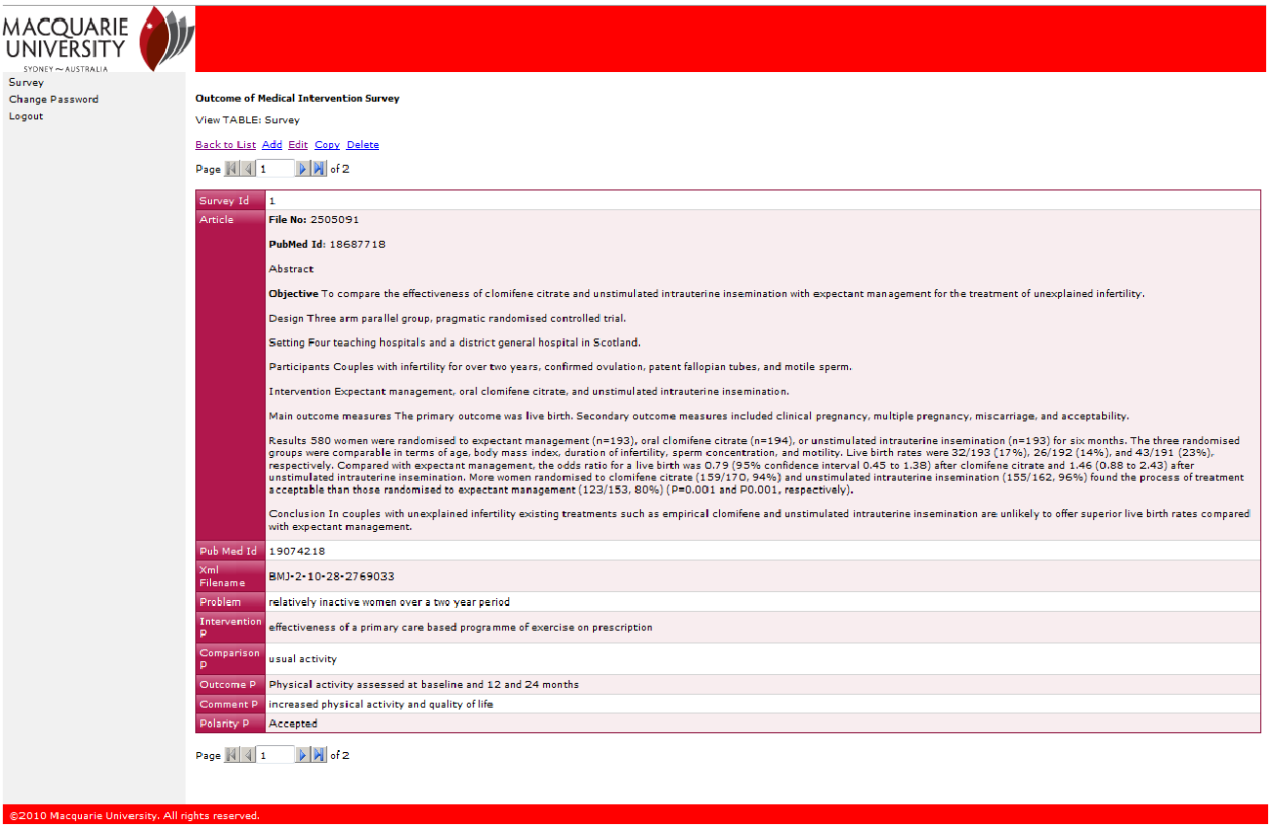

Figure 2: View annotations details page

with larger training data. Therefore we considered these classifiers only as baselines to improve. We used the following features:

1. All words in the abstract;
2. All words in the conclusion section;
3. Selected words in the abstract; and
4. Selected words in the conclusion section.

All abstracts were originally structured into sections, and, therefore, it was trivial to select the conclusion section. The selection of words was done by visual inspection of the training dataset. The following words were selected: *achieved, decrease, decreased, difference, effect, effective, effects, efficacy, improve, improvement, increase, increased, no, not, provide, provided, reduce, reduced, significant.*

The classifiers selected were decision tree (J48), support vector machine (SVM), and Naïve Bayes (NB). The results are shown in Table 2.

The best results were obtained by using the selected words in the conclusion section by the J48 classifier.

Note, however, the large confidence intervals[9] that are due to the small size of the test set (67 documents). Still, even with these large confidence intervals we observe statistically significantly better results by focusing on the conclusion section versus using the complete abstracts. The difference between using all words or only a selection was not statistically significant and it would be interesting to repeat the classification with more data to determine whether the difference would become statistically significant.

3.4 Rule-based Classifier

Our rule-based classifier was based on NegEx, which was simplified as follows:

1. A different list of negation triggers was compiled by examining the frequency of triggers in the training dataset (see below).
2. The classes were changed from "affirmed" to "Accepted" and "negated" to "Rejected". The class "possible" was removed.
3. The function that sorts the negation triggers and the function that uses regular expressions to match a set of negation triggers with text in each sentence were retained with very minor changes.
4. The functions in NegEx related to the detection and use of concept sentences, such as shortness of *breath, headache, chills, fever,* etc., were removed since we integrated the concepts related to evidence in the list of negation triggers.
5. The functions that tag conjunction and pseudo-negation were also removed.

Other minor modifications were made to make the system work with the dataset used for the experiments:

6. The output to the tagger was modified to return the PubMed report ID, the conclusion that was stripped from the abstract, the abstract itself, the current class, the tagged negation phrase when found, and the system class.

[9] Confidence intervals were based on a binomial distribution and were computed using R's `Hmisc::binconf` function.

been overestimated, cannot endorse, cannot recommend, did not reduce, does not reduce, effectiveness overestimated, failed to, ineffective in, low probability, neither altered, no advantage, no advantageous, no beneficial, no benefit, no certain, no conclusive, no convincing, no definite, no detectable, no difference, no effect, no evidence, no favourable, no findings, no important, no improved, no increase, no irrefutable, no major, no meaningful, no more, no new, no novel, no overall benefit, no overall benefits, no overall effect, no positive, no proof, no reduction, no significant, no statistically, no strong, no substantial, no suggestion, nonsignificant improvement, nonsignificant improvement, nonsignificant reduction, non-significant reduction, nor protected, not affect, not appear to, not appreciate, not associated, not be, not beneficial, not change, not clinically, not confirm, not confirmed, not demonstrate, not differ, not exhibit, not find, not had, not have, not improve, not increase, not influence, not know, not known, not lead, not lend support, not likely, not meaningful, not meaningfully, not met, not necessarily, not observed, not offer, not prevent, not produce, not promote, not prove, not provide, not result, not reveal, not see, not show, not shown, not significant, not significantly, not slow, not statistically, not superior, not suppress, not to, not,, remains unproved, similarly effective, unlikely to

Table 3: List of negation phrases

7. A function was added to parse the XML files downloaded from PubMed Central. The function extracts the text of the abstract element and writes the PMCID and also the abstract to a CSV file that was used as input for the negation program. The annotated class is first entered into this CSV file before being used as input to the negation tagger.
8. A function was added to split the text of the abstract in order to separate the conclusion from the abstract.
9. The classifier incorporated functionality to output the results into a text file in the form of PMCID, Abstract, Conclusion, Tagged Sentence, Current Class, and System Class.

The following feature in NegEx has not been modified in the present version of the negation tagger:

10. The output that calculates the accuracy by comparing the current class with the class found by the classifier.

The list of negation triggers mentioned in item 1 are mostly bigrams and a few trigrams that were extracted by manually inspecting the training data. Phrases such as *no increase, no decrease, no significant, not improve*, and *not found*, often appear in the conclusion of articles where the hypothesis has been rejected. Some trigrams such as *not lend support, no major effect, no overall effect,* and *no overall benefit* also serve to negate the outcome of the intervention. The full list is shown in Table 3.

The rule-based classifier is designed to detect negation in the conclusion section only and then classify the article accordingly. The striking improvement of

	J48	SVM	NB
Baseline 1	49% (37%-61%)	66% (54%-76%)	69% (57%-79%)
Baseline 2	82% (71%-89%)	78% (67%-86%)	71% (59%-80%)
Baseline 3	54% (42%-65%)	63% (51%-73%)	58% (46%-69%)
Baseline 4	**84% (73%-91%)**	80% (69%-88%)	78% (67%-86%)

Table 2: Accuracy of the baseline classifiers with 95% confidence intervals

results obtained by restricting the analysis to the conclusion section by the statistical classifiers led us to this decision. Limiting the scope of a negation to the conclusion reduces the likelihood of false negatives, that is negation being detected and the article classed as rejected when in fact the research hypothesis has been accepted. All abstracts were structured and therefore it was trivial to select the conclusion section.

The rule-based classifier obtained an accuracy of 95% on the test set, with a 95% confidence interval between 88% and 98%. The better performance of the rule-based classifier with respect to the baselines is encouraging, though given the small amount of data used by our statistical classifiers we cannot rule out the possibility that statistical classifiers trained with more data would outperform the rule-based classifier.

An analysis of the classification errors indicated that often the error was due to the context of the negation. Even after selecting the conclusion section only, sometimes the negation detected referred to a secondary outcome rather than the main outcome. In other cases, the expression related to the evidence appeared in the results section but not in the conclusion section, and therefore it was not detected.

4 Conclusions

We have presented a rule-based classifier that detects whether the abstract of a published clinical RCT indicates whether their research hypothesis is confirmed. We do this by adapting NegEx' negation detector to focus on the detection of expressions of negation of evidence. We have simplified the original NegEx and replaced its original patterns with specific patterns derived from manual inspection of a training set. Experiments on a disjoint test set show an accuracy between 88% and 98% within a 95% confidence interval.

These results are statistically significantly better than those of a set of baselines using statistical classifiers. The amount of training data used in the statistical classifiers is comparable to that of related methods such as the one by Niu et al. (2006), though we believe that they are still too small to draw conclusions about the comparison between rule-based and statistical methods. We therefore plan to repeat the experiments with larger volumes of data. By training on larger volumes of data we expect better results using statistical classifiers. In the process, we will experiment with more complex features to feed the classifiers, including the patterns used in our rule-based system.

We also observed that the results obtained from processing only the words in the conclusion sections of the abstracts are better than those obtained using the complete text. We are considering the possibility of applying automated sentence classification techniques to detect conclusion sentences such as those devised by Demner-Fushman et al. (2006) and Kim et al. (2011).

We also plan to extend the analysis to more varied types of studies. Even though RCTs represent a relatively large percentage of clinical studies, there are other important types of studies that should be considered, such as meta-analyses.

Further work also includes the detection of secondary outcomes in the papers. By detecting these we hope to reduce the scope of the negation expression and increase the overall accuracy results.

Finally, we plan to perform an extrinsic evaluation by integrating this research into an application system that returns relevant medical papers and ranks them by the quality of their clinical evidence. The research presented here will provide one additional item of information to this system. We will study the impact of this feature in the overall task.

Given that the list of negation triggers discovered in this study does not contain clinical terminology it is possible that this program with the current list of triggers would be useful to classify any experimental-based research paper with no or minor modifications.

Being a modification of NegEx, the classifier is written in Python and is available as open-source code.[10]

References

Aronow, D. B., Fangfang, F. and Croft, W. B. (1999), 'Ad hoc classification of radiology reports', *Journal of the American Medical Information Association* **6**(5), 393–411.

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F. and Buchanan, B. G. (2001), 'A simple algorithm for identifying negated findings and diseases in discharge summaries', *Journal of Biomedical Informatics* **34**(5), 301–310.

Demner-Fushman, D., Few, B., Hauser, S. E. and Thoma, G. (2006), 'Automatically identifying health outcome information in medline records.', *J Am Med Inform Assoc* **13**(1), 52–60.

Goryachev, A., Sordo, M., Ngo, L. and Zeng, Q. (2006), Implementation and evaluation of four different methods of negation detection, Technical report, DSG.

Gosall, N. and Gosall, G. (2009), *The Doctors' Guide to Critical Appraisal*, 2 edn, PasTest Ltd., Knutsford Cheshire.

Huang, Y. and Lowe, H. J. (2007), 'A novel hybrid approach to automated negation detection in clinical radiology reports', *Journal of the American Medical Informatics Association* **14**(3), 304–311.

Kim, S. N., Martinez, D., Cavedon, L. and Yencken, L. (2011), 'Automatic classification of sentences to support evidence based medicine', *BMC Bioinformatics* **12**(Suppl 2), S5.

[10] `http://sourceforge.net/p/clinevidence/`

Meystre, S. M. and Haug, P. J. (2005), Comparing natural language processing tools to extract medical problems from narrative text, *in* 'AMIA Annual Symposium Proceedings', pp. 525–529.

Mutalik, P. G., Deshpande, A. and Nadkarni, P. M. (2001), 'Use of general-purpose negation detection to augment concept indexing of medical documents', *Journal of the American Medical Informatics Association* **8**, 598–609.

Niu, Y., Zhu, X. and Hirst, G. (2006), Using outcome polarity in sentence extraction for medical question-answering, *in* 'Proceedings AMIA'.

Skeppstedt, M. (2010), Negation detection in swedish clinical text, *in* 'Proceedings NAACL HLT Second Louhi Workshop on Text and Data Mining of Health Documents', pp. 15–21.

Uzuner, Ö., Zhang, X. and Sibandab, T. (2009), 'Machine learning and rule-based approaches to assertion classification', *Journal of the American Medical Informatics Association* **16**, 109–115.