Artificial Intelligence in Medicine 51 (2011) 17–25

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Exploiting the systematic review protocol for classification of medical abstracts

Oana Frunza[a,*], Diana Inkpen[a], Stan Matwin[a], William Klement[a], Peter O'Blenis[b]

[a] School of Information Technology and Engineering, University of Ottawa, 800 King Edward, Ottawa, Ontario, Canada K1N 6N5
[b] Evidence Partners Corporation, 9 Wick Crescent, Ottawa, Ontario, Canada K1J 7H1

ARTICLE INFO

Article history:
Received 18 January 2008
Received in revised form
22 September 2010
Accepted 14 October 2010

Keywords:
Automatic text classification
Text representation
Medical concepts
Ensemble of classifiers
Systematic reviews for the medical domain

ABSTRACT

Objective: To determine whether the automatic classification of documents can be useful in systematic reviews on medical topics, and specifically if the performance of the automatic classification can be enhanced by using the particular protocol of questions employed by the human reviewers to create multiple classifiers.
Methods and materials: The test collection is the data used in large-scale systematic review on the topic of the dissemination strategy of health care services for elderly people. From a group of 47,274 abstracts marked by human reviewers to be included in or excluded from further screening, we randomly selected 20,000 as a training set, with the remaining 27,274 becoming a separate test set. As a machine learning algorithm we used complement naïve Bayes. We tested both a global classification method, where a single classifier is trained on instances of abstracts and their classification (i.e., included or excluded), and a novel per-question classification method that trains multiple classifiers for each abstract, exploiting the specific protocol (questions) of the systematic review. For the per-question method we tested four ways of combining the results of the classifiers trained for the individual questions. As evaluation measures, we calculated precision and recall for several settings of the two methods. It is most important not to exclude any relevant documents (i.e., to attain high recall for the class of interest) but also desirable to exclude most of the non-relevant documents (i.e., to attain high precision on the class of interest) in order to reduce human workload.
Results: For the global method, the highest recall was 67.8% and the highest precision was 37.9%. For the per-question method, the highest recall was 99.2%, and the highest precision was 63%. The human–machine workflow proposed in this paper achieved a recall value of 99.6%, and a precision value of 17.8%.
Conclusion: The per-question method that combines classifiers following the specific protocol of the review leads to better results than the global method in terms of recall. Because neither method is efficient enough to classify abstracts reliably by itself, the technology should be applied in a semi-automatic way, with a human expert still involved. When the workflow includes one human expert and the trained automatic classifier, recall improves to an acceptable level, showing that automatic classification techniques can reduce the human workload in the process of building a systematic review.

1. Introduction

Systematic reviews are highly structured summaries of existing research in a particular field. They are a valuable tool in enabling the spread of evidence-based practices especially in the medical domain as the amount of information in medical publications continues to increase at a tremendous rate. Systematic reviews help to parse this growing body of information and distill targeted knowledge from it.

The systematic review process, though typically less expensive than primary research, requires considerable time and effort, as human reviewers must screen references manually to determine their relevance to each given review. This process often entails reading thousands or even tens of thousands of article abstracts. The continuing growth of the body of medical articles makes this process increasingly difficult.

A systematic review begins with a query-based search to identify articles that may be candidates for inclusion. Two reviewers then read each abstract to determine whether the entire article (which may not be available for free) should be examined. If so, further analysis of the article decides whether it is clinically relevant to the review topic and what information should be extracted.

A systematic review must be exhaustive; the accidental exclusion of a potentially relevant abstract can have a significant negative impact on the validity of the overall review [1]. Thus the process is extremely labor-intensive.

* Corresponding author. Tel.: +1 613 562 5800x2140; fax: +1 613 562 5175.
E-mail address: ofrunza@site.uottawa.ca (O. Frunza).

0933-3657/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2010.10.005

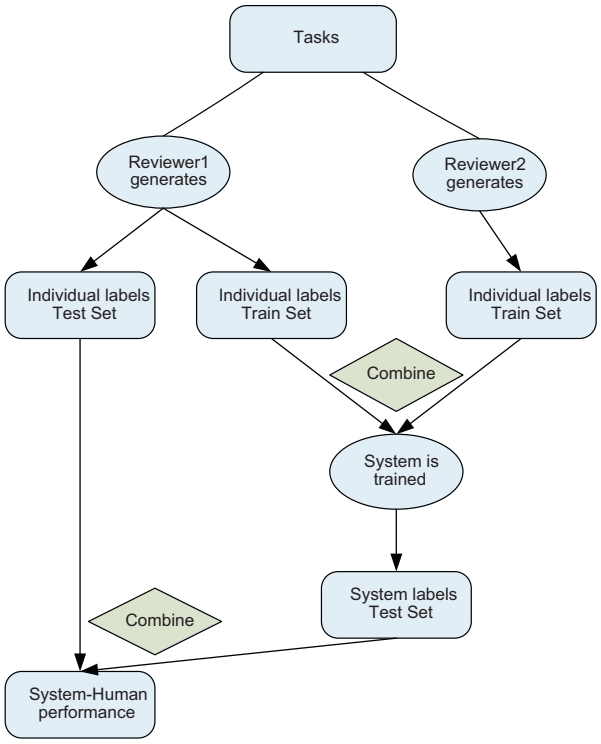

Fig. 1. Embedding automatic text classification in the process of building a systematic review.

This paper proposes using an automatic system during the initial (abstract) screening phase in order to reduce the human effort involved in preparing a systematic review. Under the proposed approach, one reviewer will still read the entire collection of abstracts, but the other reviewer will have to label only the articles that will be used to train the classifier, the rest of the articles will be labeled by the classifier. Ideally the proportion of articles that the reviewer must label in order to train the classifier will be small, so as to achieve a higher workload reduction.

We envision two ways to obtain the labels of the abstracts that will be used in training the classifier. The labels could be based only on the decisions made by the assisted reviewer, or they could represent the final decisions resulting from the work of both reviewers. Usually, if either reviewer believes that an article should receive further screening, it is labeled for inclusion (the benefit of doubt plays an important role in the decision process). The decision process for the labels when the two reviewers' opinions are used can be the same as the one used in the initial screening phase: if at least one reviewer agreed to include the abstract, the abstract will have the labeled as included. For the experiments performed in this paper, we used the labels obtained after the two reviewers' decisions are combined. This approach should both maintain reliability of the systematic review and reduce the overall workload. With regard to reliability, even if one of the reviewers is assisted by an automatic classifier, the chances that both the human judge and the classifier exclude the same abstract will be approximately the same as if two human judges had directly reviewed the abstract. The reduction in workload is from the time required for the usual two passes through the whole collection of abstracts (by both humans) to only one full pass plus a lesser amount of activity by the classifier-assisted reviewer.

Fig. 1 graphically presents in flowchart form the process of building a systematic review when the labels for training the classifier are based on the decisions of both reviewers. Alternative processes are also possible; for example, some of the abstracts labeled by the classifier could be double-checked by the assisted human reviewer who would then make the final labeling decisions.

An automatic system helping with the tedious process of deciding the relevance or non-relevance of each abstract could make systematic reviews easier, faster, more scalable, and more affordable to complete. Machine learning techniques could fulfill this need [2]. Specifically, a subfield of machine learning called automatic text categorization is highly relevant to the development of an intelligent systematic review system, since the task that must be completed is a text classification task intended to classify an abstract as relevant or not relevant to the topic of review.

The methods described in this paper apply machine learning to the preparation of systematic reviews. The hypothesis guiding this research is that replacing some of the manual screening of abstracts with the use of an automatic classifier, which can be trained to determine the relevance of abstracts at modest cost, will save time while still achieving good performance. The experiments described herein are designed to show that appropriate methodological design and classification algorithms can attain this combination of reduced effort and suitably rigorous review.

2. Background

The traditional way to collect and triage the abstracts in a systematic review begins with the use of simple query search techniques based on MeSH (www.nlm.nih.gov/mesh, accessed on 24 September 2008) or keyword terms. The queries are usually Boolean-based and are optimized either for precision (to retrieve only few non-relevant articles) or for recall (to miss as few relevant articles as possible). Studies such as [3] show that it is difficult to obtain high performance for both measures.

Although the task of selecting papers for a systematic review is a natural application of a well-developed area of automatic text classification, prior efforts to exploit this technology for such reviews has been limited. The research done by [2] appears to be the first such attempt. In that paper, the authors experimented with a variety of text classification techniques, using the data derived from the ACP Journal Club[1] as their corpus. They found that support vector machine (SVM) was the best classifier according to a variety of measures, but could not provide a comprehensive explanation as to how SVM decides whether a given abstract is relevant. The authors emphasized the difficulties related to the predominance of one class in the datasets (i.e., the number of relevant abstracts is only a small portion of the total), along with the difficulty of achieving both good recall and good precision.

Further work was done by [1], focused mostly on the elimination of non-relevant documents. As their main goal was to save work for the reviewers involved in systematic review preparation, they defined a measure, called work saved over sampling (WSS), that captured the amount of work that the reviewers would save with respect to a baseline of just sampling for a given value of recall. The idea is that a classifier can return, with high recall, a set of abstracts, and that the human needs to read only those abstracts and weed out the non-relevant ones. The savings are measured with respect to the number of abstracts that would have to be read if a random baseline classifier were used. Such a baseline corresponds to uniformly sampling a given percentage of abstracts (equal to the desired recall) from the entire set. In the work done by [1], the WSS measure was applied to report the reduction in reviewers' work when retrieving 95% of the relevant documents, but the precision was very low. The present study focuses on developing a classifier for systematic review preparation, relying on characteristics of the

[1] http://www.acpjc.org/.