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

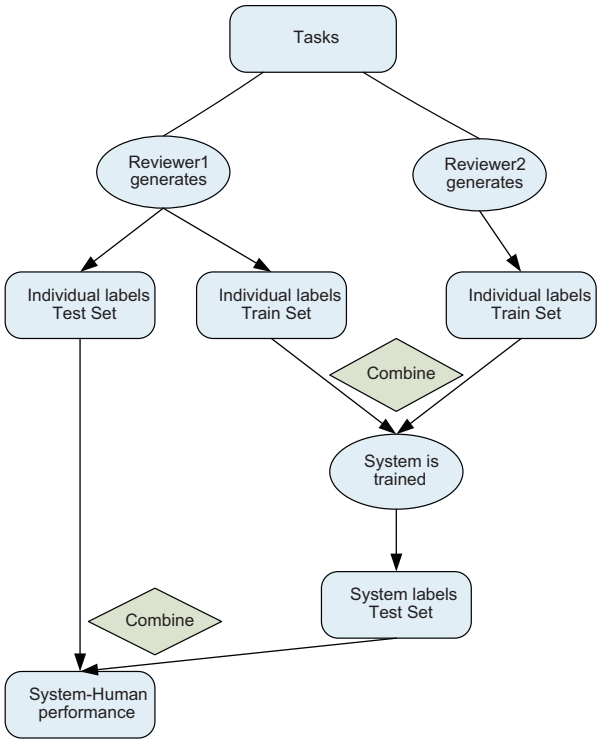

Fig. 1. Embedding automatic text classification in the process of building a systematic review.

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/.

Table 1
The set of questions used by reviewers to determine if an abstract is relevant to the systematic review's topic.

ID	Question	Type of response
1	Is this article about a dissemination strategy or a behavioral intervention?	Yes/No
2	Is this paper relevant as background although it does not meet inclusion criteria?	Yes/No
3	Is the population in this article made of individuals age 65 or older, or does it comprise individuals who serve the elderly population's health needs (i.e., health care providers, policy makers, organizations, community)?	Yes/No
4	Is this a primary study?	Yes/No
5	Is this article about:[a]	Text
6	Other reason for exclusion. Please specify	Text
7	Is this a review?	Yes/No
8	Is there an abstract? Answering this question does not create an exclusion rule.	Yes/No

[a] This is an open-ended question which the reviewers can answer by typing what they believe is the topic of the article. This information can be used later in the process of building the systematic review.

data that were not included in the method used by [1], and therefore we cannot perform a direct comparison of results here. Also, the data sets that [1] used in their experiments are significantly smaller than the one that we used here.

The following issues are central to the process of using machine learning techniques for systematic review tasks:

Distribution of the data – in most systematic reviews, the number of relevant documents is much smaller than the number of non-relevant documents creating a class imbalance which can cause problems for the machine-learning algorithms.

Noise – when reviewers are not sure from an abstract whether the article is relevant, a final decision will be made during the second screening process where the entire document is reviewed. This "benefit of the doubt" approach will affect the quality of the data used to train the classifier, since a certain amount of "noise" is introduced, in that abstracts that are in fact non-relevant are often labeled as relevant in the first screening process.

Labeling cost – the labeling process is expensive both in human effort and money, so a key goal of the machine learning approach is to reduce the human effort required.

Misclassification cost – if an abstract does not pass the first level of screening, the article will not be examined for information extraction during the second level of screening. Failure to identify a relevant abstract in the first screening process can have a profoundly negative impact on the validity of systematic review results [1]. Identifying all relevant abstracts, therefore, is critical.

Representation – due to the vast number of abstracts in the medical domain repositories, the machine learning representation must take into account the huge number of attributes or words that can be extracted.

Training cost – if the overhead cost associated with training and tuning a machine learning classifier is too high, this expense may negate the economic value of the solution. Similarly, if the training interface and process for the machine learning classifier are too complex for a non-technical user, its relevance to the systematic review community will be negligible.

3. Methods

In our current work we propose two machine learning techniques to assist humans in the systematic review process. The first one uses standard text classification techniques [1], while the second one is a new technique that takes into account the specifics of the systematic reviews – that is, the questions and answers that are commonly part of the systematic review protocol. We believe that this second approach is an especially promising way to obtain desirable results.

3.1. The data set

A set of 47,274 abstracts with titles was collected from MEDLINE (http://medline.cos.com, accessed on 24 September 2008) as part of a systematic review done by McMaster University's Evidence-Based Practice Center using TrialStat Corporation's Systematic Review System (www.trialstat.com, accessed on 24 September 2008), a web-based software platform used to conduct systematic reviews.

The initial set of abstracts was collected using a set of Boolean search queries that were run for the specific topic of the systematic review (which was "the dissemination strategy of health care services for elderly people of age 65 and over"). Normally, such queries are intentionally very general so as not to miss any relevant abstracts. The goal of the queries is to provide close to 100% recall, even if at the expense of precision. It is not known if this bound of recall performance is actually achievable. The explicit terms that created the query are not disclosed; we only have the collection of abstracts and the label attached by the reviewers to each one. The process by which labels were assigned is described below, as it follows a specific systematic review protocol. The label determines if the abstract is relevant or not to the topic of interest.

In the protocol applied, two reviewers worked in parallel, reading the entire collection of 47,274 abstracts and answering a set of questions to determine if an abstract is relevant or not to the topic of review. Table 1 presents the set of questions used to determine if an abstract was relevant or not.

The set of questions, or screening protocol, is prepared during the setup stage of a systematic review when the topic of the review and the search space are defined.

Using the questions presented above and the answers of the two reviewers, an abstract was considered not relevant in the first screening phase if at least one question was answered negatively by both reviewers. Otherwise the abstract was selected for stricter screening in the next phase.

In our experiments we worked with the collection of abstracts and labels determined from the two reviewers' opinions, after possible disagreements were solved (the two reviewers could discuss and change the labels for some abstracts before they were passed into the second level of screening). From the entire collection of abstracts, 7173 were considered relevant and the remaining 40,101 non-relevant. As noted previously, when the abstract did not enable reviewers to decide whether an article was relevant, the reviewers preferred to keep the document, giving it the benefit of the doubt, so as not to risk overlooking a relevant article. Some of the documents that passed the first screening process were in fact discarded in the second screening process, when the full-text of the article offered more information to support the decision. The abstracts considered irrelevant in the initial screening phase do not enter the next screening steps; they are removed from the systematic review. The method proposed in this article focuses on the initial screening phase, during which decisions are made based only on the text of the abstract.

Using the collection of abstracts and their titles, labeled as relevant or not relevant to the topic of the review, our task is to develop

		Predicted	
		Included	*Excluded*
Actual	*Included*	*TI*	*FE*
	Excluded	*FI*	*TE*

Fig. 2. Example of confusion matrix [TI = the number of true inclusions; FE = the number of false exclusions; FI = the number of false inclusions; TE = the number of true exclusions].

from this labeled collection an automatic classifier that will appropriately mark each abstract as relevant (included) or non-relevant (excluded).

3.2. Evaluation measures

When we evaluate performance in classifying abstracts, two objectives are of great importance: to ensure the completeness of the systematic review (i.e., to maximize the number of relevant documents included) and to reduce the reviewers' workload (by maximizing the number of irrelevant documents excluded). Our evaluation measures of machine learning system performance are computed from the confusion matrix [4] that contains information about the actual class into which each abstract falls according to human review and the class selected by the classifier. Fig. 2 presents an example of a confusion matrix.

The following evaluation measures, commonly used in information retrieval and classification tasks are used here:

Recall = the ratio of correctly classified included instances to the total number of included = TI/(TI + FE). This evaluation measure is known to the medical research community as sensitivity. It indicates the proportion of items that the system incorrectly failed to select and it measures how well we achieve our Objective 1.

Precision = the ratio of correctly classified included instances to the total number classified as included = TI/(TI + FI). This measure reflects our Objective 2 and assesses to what extent the system has inaccurately included abstracts that it should have excluded as non-relevant.

F-measure = the harmonic mean between precision and recall = 2 * Precision * Recall/(Precision + Recall) [5].

Since Objective 1 is more important than Objective 2, a decrement of recall (leaving out relevant documents) is more costly than a decrement of precision (including non-relevant documents that will be filtered out in the next screening stage).

For comparison purposes, we use the results of a simple query-based system as an extreme baseline, along with a random-baseline classifier that randomly generates a real number between 0 and 1 and uses the prevalence of the two classes to decide. If the generated number is less or equal to the threshold the label is relevant, otherwise it is non-relevant. An automatic classification system must, obviously, exceed the results attained by these two baselines to be of any value.

3.3. Global text classification method

The first method that we propose for the text classification task entailed in the systematic review process is a straightforward machine learning approach. We trained a classifier on a collection of abstracts and then evaluated the classifier's performance on a separate collection of abstracts representing the test data set. The power of this classification technique, also used by [1], stands in the ability to use a suitable classification algorithm and a good representation for the text classification task.

Table 2
Training and testing data sets.

Data set	No. of abstracts	Class distribution (included:excluded) (ratio)
Training	20,000	3,056: 16,944 (1:5.6)
Testing	27,274	4,117: 23,157 (1:5.6)

We used the same machine learning algorithm to train several classifiers for the second method that we propose, described in detail in Section 4. In this method we used additional knowledge from the systematic review protocol described in Section 3.1 with an ensemble of classifiers trained on different data sets, with a final classification decision for each testing instance taking into account the predictions of the ensemble of trained classifiers. This approach arose from consideration of the multiple, broad questions, covering different concepts, that are used in the initial screening phase of systematic reviews. It seemed that dividing up the questions (among multiple classifiers) and then combining the end results could increase performance. As we will show later in the results section, our intuition was confirmed.

We made the transition from the first method to the second method by first combining the global method with information from the answers that the reviewers gave to the questions for each abstract. This information takes the form of the three possible answers to each closed-ended question (yes, no, or not applicable). We randomly split the data set of abstracts into a training set and a test set. We used the first part of the split for training and the second one to evaluate the classifier's performance in deciding whether to include or exclude an abstract. We decided to work with a training set smaller than the test set because, ideally, good results should be obtained without relying on too large an amount of training data. We have to take into consideration that, if we want to train a classifier for a particular topic of review, human effort will be required to annotate at least part of the collection of abstracts.

From the collection of 47,274 abstracts, 20,000 were randomly taken to be part of the training data set and the remaining 27,274 represented the test set.[2] Table 2 presents a summary of the data along with the class distribution in the training and test data sets. We verified that the original overall inclusion/exclusion distribution of 1:5.6 between the two classes remains in both subsets.

To evaluate the effect of training set size, we ran additional experiments with different splits of the data set, changing its size from 10% to 50% in increments of 10%. We decided not to include more than 50% of the data for training, because doing so would increase the workload of the human judge, and our final goal is to use as little training data as possible.

3.4. Representation

The next preprocessing step to be addressed is to choose a representation of the documents that can be used in the machine learning task. We used three types of representations: bag-of-words (BOW), concepts from the Unified Medical Language System (UMLS), and a combination of BOW and UMLS concepts.

The bag-of-words representation is commonly used for text classification. We have chosen to use binary feature values, which have been shown to outperform weighted values for text classification tasks in the medical domain [1] and to provide more stable results than frequency values can provide on similar tasks [6].

[2] We believe that the cross-validation evaluation technique is not appropriate for our task since splitting the data using a time-line stamp is a realistic scenario when integrating a trainable automatic system into the human workflow. Also, cross-validation is normally used when the data set is not large enough to split it into training and test sets.

The selected features were words of greater than three characters, delimited by spaces and simple punctuation marks that appeared at least three times in the training collection. The frequency threshold of three is commonly used for text collections, especially large ones, because it removes non-informative features and strings of characters that might be the result of a wrong tokenization when splitting the text into words, and because very short acronyms in the medical domain could be highly ambiguous. We also removed stop words (e.g., *the*, *it*, *of*, *an*), using the stop-word list (www.site.uottawa.ca/~diana/csi5180/StopWords, accessed on 24 September 2008). These function words appear in every document and, therefore, do not help in the classification.

Even with these feature selection constraints, the remaining number of features is very large – approximately 30,000 word features. In order to reduce the number of features we experimented with several feature selection algorithms, but the results were not better than when using all the features. We used the InfoGain [7], chi-square [8], and bi-normal separation (BNS) [9] feature selection techniques. No improvements in results were obtained with any of these methods.

UMLS concepts: To supply a representation that provides features more general than the words in the abstracts, we also added UMLS [10] concept representations. UMLS is a knowledge source developed at the U.S. National Library of Medicine (NLM) that contains a metathesaurus, a semantic network, and the specialist lexicon for biomedical domain.

The metathesaurus is organized around concepts and meanings; it links alternative names and views of the same concept and identifies useful relationships between different concepts.

UMLS contains more than 1 million biomedical concepts and more than 5 million concept names, which are hierarchically organized. Each unique concept present in the thesaurus has multiple text string variants (slight morphological variations of the concept) associated with it. NLM created this knowledge base by unifying hundreds of other medical knowledge bases and vocabularies (such as MeSH and SNOMED CT) to create an extensive resource that provides synonymy links, as well as parent–child relationships, among single or multi-word concepts. All concepts are assigned at least one semantic type from the semantic network; this linkage provides a generalization of the existing relations between concepts. There are about 135 semantic types in the knowledge base, and they are linked through 54 relationships.

In addition to the UMLS knowledge base, NLM created a set of tools that allow easier access to the useful information. MetaMap [11], one of these tools, maps free text to biomedical concepts in the UMLS, or, equivalently, it discovers metathesaurus concepts in a text. With this software, text is processed through a series of modules that in the end will give a ranked list of all possible concept candidates for a particular noun-phrase. We used as features the top concept candidate for each phrase identified by the MetaMap system. The UMLS concept representation is similar to a multi-word expression representation. For the simple BOW representation, each feature is represented by a single word. Using UMLS concepts, features are often represented by a sequence of words.

Another reason to use a UMLS concept representation is the "concept drift" phenomenon that can appear in a BOW representation. This is an especially frequent problem in the medical domain [12]. New articles on a certain topic frequently use new terms that might not match the ones encountered previously in the training process. Using a more general representation rather than individual words can still capture these articles.

3.5. Classification algorithms

As a classification algorithm we chose to use the complement naive Bayes (CNB) [13] classifier from the Weka [4] tool. The reason for this choice is that the CNB classifier implements state-of-the-art modifications of the standard multinomial naive Bayes (MNB) classifier for a classification task with highly skewed class distribution. As the systematic reviews usually identify a large majority of abstracts as not relevant, resulting in skewness reaching even below 1%, it is appropriate to use classifiers that take this problem into account. CNB modifies the standard MNB classifier by applying asymmetric word count priors, reflecting skewed class distribution [13]. We experimented with other classifiers from Weka as well (such as decision trees, support vector machine, instance-based learning, and boosting), and the results obtained with CNB were better than those with the other classifiers.

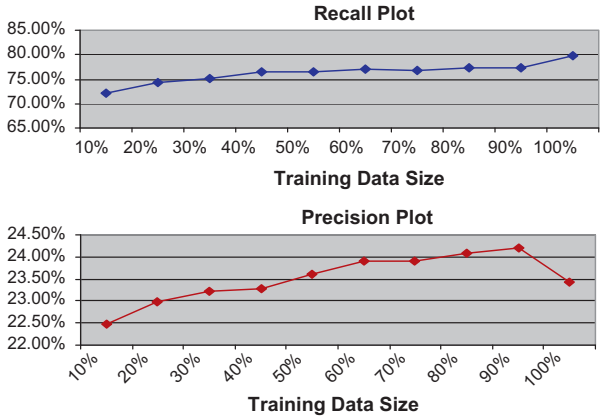

Fig. 3. Algorithm for per-question classification method. Recall and precision plots when varying the training size for per-question technique.

Naïve Bayes (NB) is a classifier that is known to work well with text. It is fast and easy to integrate in a more complex system. Its performance is comparable to that of the SVM classification, and a previous study [14] showed that the NB classification algorithm works better than SVM with a high number of features. In that study the authors compared SVM, NB, and k-nearest neighbor (kNN) classification algorithms for text classification tasks. They found that all classifiers obtained comparable results, but that NB worked best for various sizes of the training data set. The difference in results between SVM and NB was not significant, but, from the point of view of the running times and parameter settings, the NB classifier is definitely a more suitable choice than SVM for our task.

4. Per-question classification method

The second method that we propose for solving our task takes into account the specifics of the systematic review process. More exactly, it takes advantage of the set of questions that the reviewers use to decide whether an abstract is relevant. These questions are created in the design step of the systematic review, and almost all systematic reviews include such a set of questions, similar to the ones presented in Table 1. Ref. [1] represented such information in the form of a closed set of reasons for the exclusion of a specific abstract, but this information was not used by the classifier developed in their work.

In the worst case, the design of a systematic review has only one question, in which case the method will be similar to the global text classification technique that we presented earlier. The algorithm for the method that we propose is described in Fig. 3.

We have chosen to use only the questions that have inclusion/exclusion criteria, because they are the most important ones for reviewers as they make a decision on an abstract.

To collect training data for each question, we used the same training data set as in the previous method (note that not all the abstracts have answers for all the questions; therefore the training

Table 3
Data sets for the per-question classification method.

Question	Training data set	Included class	Excluded class
Q1 – Is this article about a dissemination strategy or a behavioral intervention?	14,057	1,145	12,912
Q3 – Is the population in this article made of individuals age 65 or older, or does it comprise individuals who serve the elderly population's health needs (i.e. health care providers, policy makers, organizations, community)?	15,005	7,360	7,645
Q4 – Is this a primary study?	8,825	6,895	1,930
Q7 – Is this a review?	6,429	5,640	789

set sizes differ for each question). We also kept the same test data set. When we created a training data set for each question separately, we removed the abstracts for which only one of the human experts had given a "yes" answer, to eliminate "noise" in the training data. We need to train classifiers only on reliable data, to the extent possible. Table 3 presents the subset of questions that we selected and the number of instances (abstracts) present in each training data set for each class.

Combining the per-question classifiers: For each of the questions from Table 3, we trained a CNB classifier on the corresponding data set. We used the same representations for the per-question classifiers as we did for the global classifier: BOW, UMLS (using the concepts that appeared only in the new question-oriented training data sets), and the combination BOW + UMLS. We used each trained model to obtain a prediction for each instance from the test set; therefore each test instance was assigned four prediction values of 0 or 1. The predictions have values of 0 or 1. To assign a final class for each test instance from the prediction of all four classifiers, the class of a test instance was decided according to one of the following four schemes:

1. If any one vote is *Excluded*, the final class of a test instance is *Excluded*.
2. If any two votes are *Excluded*, the final class of a test instance is *Excluded*.
3. If any three votes are *Excluded*, the final class of a test instance is *Excluded*.
4. If all four votes are *Excluded*, the final class of a test instance is *Excluded*.

These alternatives will be referred to in the results section as the "one-vote scheme," "two-vote scheme," and so forth. When we combined votes for all of the classifiers, we gave each classifier an equal importance, to decide the final classification.

5. Results

Before presenting the results of the two methods, we mention preliminary results obtained for the purpose of considering what is the best training–testing split for the global method. These results are presented in Table 4.

These results led us to use a 40–60% split in the remaining experiments, as a 50–50 split produced only modest improvement in recall along with a loss of precision.

For the experiment in which we combined the global method with the answers for the questions involved in the systematic review protocol, we obtained a recall value of 67.5%, a precision value of 37.9%, and an *F*-measure value of 48.5%. Comparing these results with the ones that we obtained with the global method, we observed that the precision result is fairly strong but that the recall is not acceptable.

In order to determine the performance of the human–machine workflow that we propose in this paper, we computed the recall values when the human reviewer's labels are combined with the labels obtained from the proposed classifier. We present these results in Table 5. The same labeling technique is applied as for the human–human workflow: if at least one decision for an abstract is to include it in the systematic review, then the final label is *Included*.

We also calculated the evaluation measures for the two reviewers independently. The evaluation measures for the human judge who remained in the human–machine workflow, identified as Reviewer 1 in Fig. 1, are 64.29% for recall and 75.20% for precision. The evaluation measures for the reviewer to be replaced in the human–machine classification (Reviewer 2 in Fig. 1) were 59.66% for recall and 85.09% for precision. The recall value for the two human judges combined is 85.26% and the precision value is 85%. (These figures, incidentally, show the importance of using two reviewers, as the results when both reviewers participate are much higher than the results for each of them individually.)

5.1. Results for the global method

The results for the global text classification method are reported for the BOW representation, the UMLS representation, and their combination using the CNB classifier (see Table 5).

The BOW features were identified following the guidelines presented in Section 3.4, and a total of 23,906 features were selected. To determine the UMLS concepts we used the MetaMap system described earlier. From the whole training abstracts collection, a total of 459 UMLS features were identified. As Table 5 shows, under the global method the UMLS representation obtained the highest recall results, 67.8% but much lower precision, 23.8% compared to 34.9% for BOW representation. The hybrid representation BOW + UMLS had similar results to those with BOW alone. Recall increased slightly for the hybrid representation compared to BOW alone, but the level of recall is still not acceptable. These recall results indicate that it is not viable to use a single classifier in place of a human judge in the workflow of building a systematic review.

5.2. Results for the per-question classification method

We now turn to our second method, which trains a classifier for each question in the systematic review protocol. The results

Table 4
Results for the global method using the BOW representation technique for various splits of the training and test data sets.

Train-test split	True included	False included	True excluded	False excluded	Recall	Precision	*F*-measure
10–90%	1,137	1,179	34,898	5,332	17.58%	49.09%	25.89%
20–80%	2,721	3,665	28,404	3,029	47.32%	42.61%	44.84%
30–70%	2,969	5,087	23,019	2,017	59.55%	36.85%	45.53%
40–60%	2,692	5,022	18,135	1,425	65.39%	34.90%	45.51%
50–50%	2,316	4,524	14,615	1,092	67.96%	33.86%	45.20%

Table 5
Experimental results for several methods of screening 27,274 PubMed abstracts, selected for a systematic review on the dissemination strategy of health care services for people age 65 and over, of which 4117 are labeled Inc. (*Included* class) and 23,157 are labeled Exc. (*Excluded* class). For each method, the table reports the four entries of the confusion matrix (first four columns) followed by the recall and precision for the class of interest (included), the F-measure, and then the recall and precision for the human–machine workflow.

Method	True Inc.	False Inc.	True Exc.	False Exc.	Recall (%)	Precision (%)	*F*-measure (%)	Human–machine recall/precision (%)
1. Baseline methods								
IncludeAll	4117	23,157	0	0	100.0	15.0	26.2	100.0/15.0
ExcludeAll	0	0	23,157	4117	0.0	100.0	0.0	64.2/15.2
Random-Bias	370	2019	21,138	3747	8.9	15.4	11.3	67.8/15.3
2. Global method								
BOW	2692	5022	18,135	1425	65.3	34.9	45.5	87.7/17.8
UMLS	2793	8922	14,235	1324	67.8	23.8	35.2	88.6/16.9
BOW + UMLS	2715	5086	18,071	1402	65.9	34.8	45.5	87.7/17.8
3. One-vote threshold for the per-question method								
BOW	1262	745	22,412	2855	30.6	62.8	41.2	75.3/17.1
UMLS	1222	2266	20,891	2895	29.6	35.0	32.1	74.8/16.5
BOW + UMLS	1264	741	22,416	2853	30.7	63.0	41.2	75.4/17.1
4. Two-vote threshold for the per-question method								
BOW	3181	9976	13,181	936	77.2	24.1	36.8	91.6/17.0
UMLS	2603	9505	13,652	1514	63.2	21.5	32.0	86.6/16.4
BOW + UMLS	3283	10,720	12,437	834	79.7	23.4	36.2	92.7/17.0
5. Three-vote threshold for the per-question method								
BOW	3898	18,915	4242	219	94.6	17.0	28.9	97.9/15.7
UMLS	3480	16,472	6685	637	84.5	17.4	28.9	94.2/15.7
BOW + UMLS	3890	18,881	4276	227	94.4	17.0	28.9	97.8/15.7
6. Four-vote threshold for the per-question method								
BOW	4085	21,946	1211	32	99.2	15.6	27.1	99.6/15.3
UMLS	3947	20,869	2288	170	95.8	15.9	27.2	98.3/15.3
BOW + UMLS	4086	21,964	1193	31	99.2	15.6	27.0	99.6/15.3

obtained with the one-vote scheme, two-vote scheme, three-vote scheme, and four-vote scheme are presented in Table 5.

For the per-question method, we also ran experiments with various data sizes. The classifier used for these experiments is the two-vote scheme with BOW + UMLS feature representation (the classifier that obtained the best performance level). Fig. 3 presents the recall and precision plots for these experiments. The results show that precision and recall tend to increase with larger training sets. The improvement was gradual when more than 40% of the training set is used, but a jump in recall occurred when we used the full collection of abstracts in the training set.

The results obtained by this method, especially the ones using the two-vote scheme, are the best obtained so far in terms of the balance between the two objectives of recall and precision. A large number of abstracts are appropriately excluded, while very few abstracts are inaccurately excluded (far fewer than with the global classification method).

The two-vote scheme performs better than the one-vote scheme, as it eliminates the errors that can arise from a single question-based classifier. When the classifiers for two different questions agree that the abstract is relevant, the chance of correct prediction is higher. When the classifiers for three or four questions are used, precision decreases.

For the per-question technique, the recall value peaked at 99.2% with the four-vote scheme using BOW and BOW + UMLS representation techniques and CNB as classifier. However, the lowest precision value of any per-question technique, 15.6%, is obtained with the same experimental setting. It is important to aim for a high recall, but not to dismiss the precision values. A difference of even <2% in precision values can require reviewers to read additional thousands of documents, as observed in the confusion matrices for the two-vote, three-vote, and four-vote methods in Table 5. The difference in precision between 24.1% (with the two-vote method) and 15.6% (with the four-vote method) would require reviewers to go through 11,988 additional documents! The two-vote scheme has a 79% recall level, compared to 92% for the four-vote scheme; this different would represent 904 more articles missed by the two-vote classifier. However, keep in mind that recall will be higher in the human–machine workflow that we propose, because most of the articles misclassified by the machine will still be identified as relevant by the human judge. We observe this impact when looking at the recall levels for the human–machine protocol in Table 5. Keeping one human in the loop makes our method acceptable as a workflow for building a systematic review.

Taking these facts into consideration, we conclude that the best experimental setting is the two-vote scheme with BOW + UMLS features.

Fig. 4 offers a graphical comparison of the results of the various methods. The two dimensions in the figure illustrate recall and precision results for several variants of the two methods. The points have different shapes and corresponding labels in this two-dimensional space.

An ideal solution for our task would be a point in the upper right corner, combining good recall with good precision. As the figure shows, the best results are achieved by the per-question classification method. This setting delivered the best balance between the

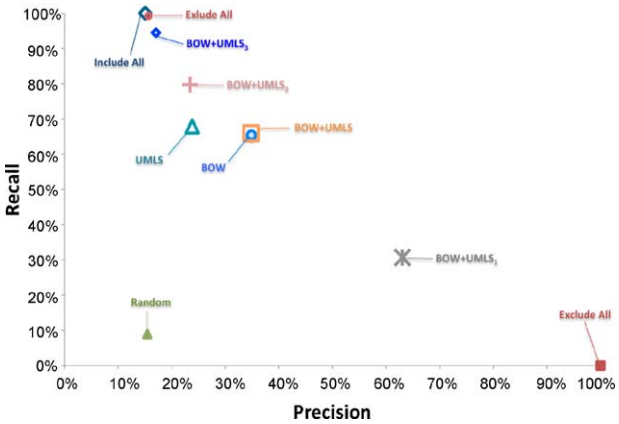

Fig. 4. Summary of results for both the global and per-question method (for the per-question method, the voting scheme used is indicated in the names by subscript numbers).

two measures, while still providing a high level of recall (the most important evaluation measure).

6. Discussion

The global method achieved good results in terms of precision, but its recall level was low and significantly outperformed by the per-question method. Overall, the best results were obtained by using the per-question method with the 2-vote scheme, including BOW representation with or without UMLS features. The results obtained by the three-vote scheme UMLS representation are close to the results obtained by the two-vote scheme, but *F*-measure results indicate that the 2-vote scheme is superior. Other per-question settings obtained better levels of recall (see the upper left corner of Fig. 4), but the level of precision is too low.

The per-question technique is also more robust, offering several options as to the desired type of performance. If the reviewers are willing to read almost the entire collection of documents in order to obtain high recall then a three- or four-vote scheme can be set up (though these schemes are not likely to achieve 100% recall because few abstracts contain answers to three or more of the questions associated with the systematic review). If the reviewers prefer to read a smaller collection of articles, nearly all of which will be relevant, then a one-vote scheme can be applied. The per-question method results confirm that a committee or ensemble of classifiers is better than one classifier; this conclusion is supported in the machine learning literature [15].

Balancing recall and precision is a known problem in the machine learning community. As our primary task is to obtain greater levels of recall, the choice of using the entire collection of training data for the per-question technique appears to be supported by the results presented in Fig. 4.

When we combined the automatic classification system with one human reviewer, we obtained a major improvement in recall. Looking at the human–machine results for the classifier alternative that obtained the best results (the two-vote scheme with a BOW + UMLS representation technique), we find that, among the 834 abstracts falsely excluded by the automatic system, 537 were identified as relevant by the first human reviewer, the one who read and labeled the entire collection of retrieved abstracts. The overall recall of the human–machine combination was 92.7%, while the precision level was 17%. When we combined the human and classifier decisions, the precision level decreased slightly relative to the one obtained by the machine only, suggesting that the first human reviewer included some abstracts that were rightly excluded by the machine.

Our goal of improving the recall level at the first level of screening was achieved, since integrating both the classifier and the human judge in the workflow enabled the recall level to jump to 92.7%. As stated in previous studies by our research group [16], the level of recall achieved by two human judges is usually between 90 and 95 percent. We can conclude that the two human judges involved in the process of building the review generally correct each other's mistakes and, together, identify nearly all relevant articles in the first screening stage. The low levels of precision are corrected in the second level of screening, where the complete documents are reviewed and more refined questions are asked so that only the truly relevant articles are retained.

The relatively low precision levels achieved in the first screening phase are not our primary concern. Our main objective in this phase is to achieve high levels of recall, since an abstract labeled as non-relevant at this point will not be reviewed afterwards and will be excluded from the systematic review. Precision can be achieved in the second screening phase, where the information available to make the decision is richer. Our results show that using a human–machine workflow in the first screening phase can significantly reduce the human workload in this tedious process while maintaining acceptable recall.

Further research is required before replacing a human reviewer with an automatic classifier becomes a standard practice, but the results that we obtained with the per-question method encourage us to believe that the human–machine workflow can be effectively implemented. This application of a machine learning solution could be of significant value in completing a task that human experts find both tedious and, as demonstrated by the systematic reviews released each year (e.g., the Cochrane Collaboration www.cochrane.org/index.htm, accessed on 24 September 2008), challenging.

7. Conclusions and future work

We presented two methods that can be used in automatic classification of abstracts for a systematic review task. The two methods are (1) the global method, where a classifier is built based on the labeled training set of relevant and non-relevant abstracts, and (2) an alternative method, exploiting a protocol commonly used by human reviewers in screening abstracts for systematic reviews. This protocol consists of asking a series of simple questions for each abstract and aggregating the answers to these questions into a binary decision. The skewed class distribution was taken into account by using special classifiers (CNB). When we used the per-question technique, noise in the data was reduced by removing unreliable instances where possible. Our evaluation measures paid attention to the relative importance of the two types of misclassification (with false exclusions posing a more serious problem than false inclusions).

Using a real-life data set, we have shown that applying the question-based systematic-review protocol and machine learning techniques can help us build better automatic models and thereby reduce the tedious workload for human experts. The continued use of one human judge enhances reliability, as the abstracts included by the human judge will still be examined in the next screening phase even if the automatic system overlooked them, and makes our proposed system a viable alternative to the usual two-human process of building systematic reviews.

In the future we plan to investigate whether the CNB classifier that we used in our approach, and which is designed specifically to address classification tasks with skewed class distribution, would also work well with the data sets used in [1] and possibly improve their results. Furthermore, we will focus on reducing the amount of training data required, by using active learning techniques to select the best cases for training. The system will ask a human reviewer to label a small amount of training data that would be most useful in training a classifier. Another direction of future work is to use classifiers in updating existing systematic reviews, as a classifier can be deployed to identify newly published articles from Medline that are relevant to the topic of the systematic review. In this way, updated information can be added to an already existing review on a certain topic.

Other possible directions for future research could include experiments that give variable weight to the decision of each classifier in the per-question method, or that combine the classifiers' decisions by using a meta-classifier instead of a voting technique.

Finally, it would be interesting to extend the machine learning approach to the next stage of the systematic review process, in which the full texts of the papers are screened for relevance. Since the full text would provide a much richer BOW representation the results of the automatic classification system should be significantly stronger than in the first screening phase.

Acknowledgments

This work was funded in part by the Ontario Centres of Excellence in partnership with TrialStat, and by the Natural Sciences and Engineering Research Council of Canada. The authors thank the McMaster Evidence-Based Practice Center for providing the data for use in this paper. The original project was funded under CIHR grant number KT62363 (PI:Raina P). The authors also want to thank Dr. Thomas Tran for his insightful comments and suggestions of improvement of the paper, and Yimin Ma for providing the software for selecting features using the bi-normal separation (BNS) method.

References

[1] Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. Journal of the American Medical Informatics Association 2006;13:206–19.
[2] Aphinyanaphongs Y, Aliferis C. Text categorization models for retrieval of high quality articles. Journal of the American Medical Informatics Association 2005;12:207–16.
[3] Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. Journal of the American Medical Informatics Association 1994;1:447–58.
[4] Kohavi R, Provost F. Glossary of terms. Machine Learning 1998;30:271–4.
[5] Witten IH, Frank E. Data mining: practical machine learning tools and techniques, 2nd ed. Morgan Kaufmann, San Francisco; 2005. www.cs.waikato.ac.nz/machine learning/weka/ [accessed 24.09.08].
[6] Ma Y. Text classification on imbalanced data: application to systematic reviews automation, M.Sc. thesis, University of Ottawa; 2007.
[7] Quinlan JR. Induction on decision trees. Machine Learning 1986;1:81–106.
[8] Yang Y, Pedersen J. A comparative study on feature selection in text categorization. In: ICML '97: proceedings of the fourteenth international conference on machine learning. 1997. p. 412–20.
[9] Forman G. Choose your words carefully: an empirical study of feature selection metrics for text classification. In: Proceedings of the 6th European conference on principals of data mining and knowledge discovery. London, England: Springer-Verlag; 2002. p. 150–62.
[10] National Library of Medicine, Unified medical language system fact sheet. Available at www.nlm.nih.gov/pubs/factsheets/umls.html [accessed 24.09.08].
[11] National Library of Medicine. MetaMap. Available at: http://mmtx.nlm.nih.gov [accessed 24.09.08].
[12] Cohen AM, Hersh WR, Bhupatiraju RT. Feature generation, feature selection, classifiers, and conceptual drift for biomedical document triage. In: Proceedings of the thirteenth text retrieval conference. Gaithersburg, MD: National Institute of Standards and Technology; 2004.
[13] Frank E, Bouckaert RR. Naive Bayes for text classification with unbalanced classes. In: Proceedings of the 10th European conference on principles and practice of knowledge discovery in databases. Berlin: Springer; 2006. p. 503–10.
[14] Colas F, Brazdil P. Comparison of SVM and some older classification algorithms. Text classification tasks: artificial intelligence in theory and practice, 217. Boston: Springer; 2006. p. 169–78.
[15] Dietterich T. Machine-learning research: four current directions. Artificial Intelligence 1997;18:97–136.
[16] Kouznetsov A, Matwin S, Inkpen D, Razavi A, Frunza O, Sehatkar M, et al. Classifying biomedical abstracts using committees of classifiers and collective ranking techniques. In: The 22nd Canadian conference on artificial intelligence. Berlin: Springer; 2009. p. 224–8.