Artificial Intelligence in Medicine 64 (2015) 89–103

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Automatic evidence quality prediction to support evidence-based decision making

Abeed Sarker[a,*], Diego Mollá[a], Cécile Paris[b]

[a] Department of Computing, Macquarie University, Sydney, NSW 2109, Australia
[b] Commonwealth Scientific and Industrial Research Organisation, Crn Vimiera and Pembroke Roads, Marsfield, NSW 2122, Australia

ARTICLE INFO

Article history:
Received 3 July 2014
Received in revised form 31 March 2015
Accepted 15 April 2015

Keywords:
Automatic text classification
Automatic medical evidence classification
Decision support system
Medical natural language processing
Evidence-based medicine

ABSTRACT

Background: Evidence-based medicine practice requires practitioners to obtain the best available medical evidence, and appraise the quality of the evidence when making clinical decisions. Primarily due to the plethora of electronically available data from the medical literature, the manual appraisal of the quality of evidence is a time-consuming process. We present a fully automatic approach for predicting the quality of medical evidence in order to aid practitioners at point-of-care.

Methods: Our approach extracts relevant information from medical article abstracts and utilises data from a specialised corpus to apply supervised machine learning for the prediction of the quality grades. Following an in-depth analysis of the usefulness of features (*e.g.*, publication types of articles), they are extracted from the text via rule-based approaches and from the meta-data associated with the articles, and then applied in the supervised classification model. We propose the use of a highly scalable and portable approach using a sequence of high precision classifiers, and introduce a simple evaluation metric called average error distance (AED) that simplifies the comparison of systems. We also perform elaborate human evaluations to compare the performance of our system against human judgments.

Results: We test and evaluate our approaches on a publicly available, specialised, annotated corpus containing 1132 evidence-based recommendations. Our rule-based approach performs exceptionally well at the automatic extraction of publication types of articles, with *F*-scores of up to 0.99 for high-quality publication types. For evidence quality classification, our approach obtains an accuracy of 63.84% and an AED of 0.271. The human evaluations show that the performance of our system, in terms of AED and accuracy, is comparable to the performance of humans on the same data.

Conclusions: The experiments suggest that our structured text classification framework achieves evaluation results comparable to those of human performance. Our overall classification approach and evaluation technique are also highly portable and can be used for various evidence grading scales.

1. Introduction

Evidence-based medicine (EBM) is a practice that requires medical practitioners to obtain the best quality clinical evidence from published research when answering clinical queries, in addition to using their own expertise. It has been described as "*the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients*" [1]. To use the best available medical evidence for solving patients' problems, practitioners are required to perform a number of steps including searching for evidence, selecting the best available evidence, extracting relevant information, and appraising the quality of the extracted evidence in the light of the patients' problems. Currently, the process of evidence-based answer generation is a manual process and primarily due to the plethora of electronically available medical documents, practitioners generally face the problem of information overload. Research has shown that practitioners often fail to pursue evidence-based answers to their clinical queries, particularly at point-of-care, due to time constraints [2]. The time associated with seeking and appraising information is largely considered to be the biggest obstacle in EBM practice [3–10]. As such, approaches that can extract relevant information from medical text, and utilise them to automatically perform some of the tasks associated with

* Corresponding author at: Department of Biomedical Informatics, Arizona State University, 13212 East Shea Boulevard, Scottsdale, AZ 85259, USA.
Tel.: +1 480 884 0349.
E-mail address: abeed.sarker@asu.edu (A. Sarker).

http://dx.doi.org/10.1016/j.artmed.2015.04.001
0933-3657/© 2015 Elsevier B.V. All rights reserved.

evidence-based decision making, can significantly aid the practice.

The appraisal of the quality of the extracted evidence is a crucial task in the process of evidence-based answer generation, and its purpose is to indicate the reliability of the recommendations that are made based on the available evidence. The quality of the best available evidence may depend on a large number of factors. For example, it may depend on the topic. The reliability of the evidence associated with different topics may vary depending on the amount of research the topics have received. Topics that have received more research attention in the past are likely to contain better quality evidence (*e.g.*, *safe behavioural interventions for obesity*), compared to topics that have received little (*e.g.*, *duration of steroid therapy for contact dermatitis*). Also, sometimes findings from different studies are not consistent, making the evidence unreliable. When making evidence-based recommendations, practitioners have to take these and other factors into account in order to assess the reliability of the extracted evidence. Thus, when extracting evidence from medical publications regarding a topic, practitioners also have to spend significant amounts of time to appraise the quality of the evidence associated with the topic.

In this paper, we describe an approach to automate the process of appraising the quality of the evidence. Our approach attempts to extract relevant information from medical abstract texts and the associated meta-data, and utilise the information to predict the quality of the evidence presented by the data. We apply natural language processing (NLP) techniques to extract features from the texts, and use the features in a supervised machine learning model to perform the quality predictions. Using a corpus that specialises in EBM question answering, we first perform an analysis of the features that are likely to be indicative of the quality of evidence. Following the analysis and the selection of the features, we apply a sequential classification model to automatically predict the quality of evidence on a discrete scale. Our approach achieves an accuracy of 62.84% when evaluated against a gold standard. Our evaluations also show that the difference between the performance of our system and that of human experts on the same data is not statistically significant.

The rest of the paper is organised as follows. We provide background on evidence appraisal including a discussion of the discrete scale that we use, and discuss some related research in Section 2. In Section 3, we discuss the data, our preliminary analysis of features, the fully automatic grade classification model, and our human evaluation experiments. In Section 4, we present the results of all our experiments along with discussions of the results. We conclude the paper in Section 5.

2. Background and related work

Due to the importance of appraising and specifying the quality of evidence in EBM practice, standardised grading scales have been proposed in the literature. Various organisations and publications have their own measure of evidence and, according to a research report produced by the Agency of Healthcare Research and Quality [11], more than 100 evidence grading scales are in use today. The report also proposes that any system for grading the strength of evidence should consider three key elements: quality (the extent to which the identified studies minimise the opportunity for bias), quantity (the number of studies and subjects included in those studies) and consistency (the extent to which findings are similar between different studies on the same topic). Among other requirements, studies have specified the need for a balance between simplicity (such that assessing the quality of evidence is not very time-consuming) and clarity (so that evidence can be easily classified into a specific grade) [12]. Comprehensiveness of grading systems is also seen as an important factor [13] since they need to be applied to studies of screening, diagnosis, prevention, therapy and prognosis. Based on these requirements, we chose the strength of recommendation taxonomy (SORT) [13] as our target grading scale. SORT was designed to provide a uniform recommendation-rating system that could be applied throughout the medicine literature. It is simple and straightforward, and, therefore, easy for practitioners to use during everyday practice. This taxonomy uses only three ratings – A (strong), B (moderate) and C (weak) – to specify the *strength of recommendation* of a body of evidence. Furthermore, the availability of a specialised corpus [14] that uses SORT as the target scale for quality prediction/grading makes this scale an ideal choice for our research. The corpus, described in the next section, enables us to compare the automatically generated evidence grades to grades assigned by human experts, and evaluate the performance of our system.

Research related to ours has focused mostly on text classification in the medical domain and automatic quality assessment of medical publications. Text classification techniques have been applied to clinical text of various granularities (*e.g.*, abstracts, sentences, phrases, and so on), from various types of sources (*e.g.*, scientific articles, clinical notes, electronic health records, clinical free texts, and so on), and with various intents (*e.g.*, quality assessment, content categorisation, polarity classification, entity recognition, and so on) [15–21]. For purposes such as retrieval and post-retrieval re-ranking, approaches based on word co-occurrences [22] and bibliometrics [23] have been proposed for improving the retrieval of medical documents. These approaches, however, do not integrate evidence-based recommendations for appraisal. Tang et al. [24] propose a post-retrieval re-ranking approach that attempts to re-rank results returned by a search engine. Their approach is only tested in a specific sub-domain (*i.e.*, Depression) of the medical domain. Kilicoglu et al. [25] focus on identifying high quality medical articles and build on the work by Aphinyanaphongs et al. [26]. They apply machine learning and obtain 73.7% precision and 61.5% recall. More recently, Kim et al. [27] proposed the use of support vector machine (SVM) classifiers to identify high-quality systematic reviews to help EBM practitioners choose the best quality evidence. A similar classification approach has also been suggested by Adeva et al. [28] to support the creation of systematic reviews. These approaches and related research generally model the problem of quality assessment as a binary classification task, where each article may either be of *good* or *bad* quality. Also, the approaches are suitable for ranking single documents only. Our research has two primary differences with existing research on automatic quality assessment: (i) we use a more standardised and specialised scale, with the intent of automatically recommending evidence-based grades; (ii) our approach is for *bodies* of evidence, which may be single documents or multiple documents on the same topic. In our work, we experiment with some of the features that are suggested to be useful by the SORT guidelines (*e.g.*, publication types of articles), and some features that have been utilised in the past to make quality estimates (*e.g.*, journal names, publication dates) in related literature.

Ebell et al. [13] suggest that the publication types of medical articles are good indicators of their qualities. Literature in the medical domain consists of a large number of publication types such as randomised controlled trials, systematic reviews, cohort studies, case studies and so on.[1] These publication types are of varying qualities (*e.g.*, a randomised controlled trial is often of much higher quality than a case study of a single patient). Greenhalgh [29]

[1] A list of publication types used by the U.S. National Library of Medicine can be found at http://www.nlm.nih.gov/mesh/pubtypes2006.html. This list is not exhaustive [accessed 10.11.14].