Towards Automatic Grading of Evidence

Abeed Sarker[1], Diego Mollá-Aliod[1], and Cécile Paris[2]

[1] Centre for Language Technology
Department of Computing, Macquarie University
Sydney, NSW 2109, Australia
{abeed.sarker,diego.molla-aliod}@mq.edu.au,
URL: http://www.clt.mq.edu.au
[2] CSIRO – ICT Centre,
Locked Bag 17, North Ryde, Sydney, NSW 1670, Australia
cecile.paris@csiro.au
URL: http://www.csiro.au

Abstract. The practice of Evidence Based Medicine requires practitioners to extract evidence from published medical literature and grade the extracted evidence in terms of quality. With the goal of automating the time-consuming grading process, we assess the effects of a number of factors on the grading of the evidence. The factors include the publication types of individual articles, publication years, journal information and article titles. We model the evidence grading problem as a supervised classification problem and show, using several machine learning algorithms, that the use of publication types alone as features gives an accuracy close to 70%. We also show that the other factors do not have any notable effects on the evidence grades.

1 Introduction

An important step for physicians who practise Evidence Based Medicine (EBM) is the grading of the quality of the clinical evidence present in the medical literature. Evidence grading is a manual process, and the time required to perform it adds to the already time-consuming nature of EBM practice [6, 5]. The aim of our work is to identify the extent to which evidence grades can be automatically determined from specific information about each publication, such as the publication type, year of publication, journal name and title. In the following sections, we present a brief overview of EBM and evidence grading, followed by a discussion of our approach, results and planned future work towards building an automatic evidence grading system.

2 Evidence Based Medicine and Evidence Grading

EBM is the '*conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients*' [16]. Current clinical guidelines urge physicians to practise EBM when providing care for their

patients. Good practice of EBM requires practitioners to search for the best quality evidence, synthesise collected information and grade the quality of the evidence.

2.1 The Strength of Recommendation Taxonomy

There are over 100 grading scales to specify grades of evidence in use today. The Strength of Recommendation Taxonomy (SORT) [4] is one such grading scale. It is a simple, straightforward and comprehensive grading system that can be applied throughout the medical literature. Consequently, it is used by various family medicine and primary care journals, such as the Journal of Family Practice (JFP)[3]. SORT uses three ratings — **A** (strong), **B** (moderate) and **C** (weak) — to specify the Strength of Recommendation (SOR) of a body of evidence. Due to the popularity of this grading system, we use it as our target grading scheme.

2.2 Factors Influencing Evidence Grades

A number of factors influence the final grade assigned to an evidence obtained from one or more published studies. According to Ebell et al. [4], these factors include: qualities of evidence of the individual studies, types of evidence presented in the studies (i.e., patient vs disease-oriented[4]) and consistency of outcomes presented. In SORT, grade **A** reflects a recommendation based on *consistent* and *good-quality*, *patient-oriented* evidence; grade **B** reflects a recommendation based on *inconsistent* or *limited-quality patient-oriented* evidence; and grade **C** reflects a recommendation based on consensus, usual practice, opinion or *disease-oriented* evidence.

3 Related Work

To the best of our knowledge, there is no existing work addressing automatic evidence grading directly, although there is work on related topics. Related research has focused mostly on automatic quality assessment of medical publications for purposes such as retrieval and post-retrieval re-ranking, where approaches based on word co-occurrences [7] and bibliometrics [14] have been proposed for improving the retrieval of medical documents. Tang et al. [18] propose a post-retrieval re-ranking approach that attempts to re-rank results returned by a search engine, which may or may not be published research work. However, their approach is only tested in a specific sub-domain (i.e., Depression) of the medical domain. Kilicoglu et al. [9] focus on identifying high-quality medical articles and build on

[3] http://www.jfponline.com

[4] Patient-oriented evidence measures outcomes that matter to patients: morbidity, mortality, symptom improvement, cost reduction and quality of life; disease-oriented evidence measures intermediate, physiologic, or surrogate end points that may or may not reflect improvements in patient outcomes, e.g., blood pressure.

the work by Aphinyanaphongs et al. [1]. They use machine learning and obtain 73.7% precision and 61.5% recall. These approaches rely heavily on meta-data associated with the articles, making them dependent on the database from which the articles are retrieved. Hence, these approaches would not work on publications that do not have associated meta-data.

The definitions of 'good-quality evidence' [4] 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 of varying qualities[5]. For example, a randomised controlled trial is of much higher quality than a case study of a single patient. Evidence obtained from the former is thus more reliable. Greenhalgh [8] mentions some other factors that influence the grade of an evidence, such as the number of subjects included in a study and the mechanism by which subjects are allocated (e.g., randomisation/ no randomisation), but the latter is generally specified by the publication type (e.g., randomised controlled trial) of the article. Recently, Sarker and Mollá [17] emphasised on the importance of publication types for SOR determination and showed that automatic identification of high-quality publication types (e.g., Systematic Review and Randomised Controlled Trial) is relatively simple. Lin and Demner-Fushman [3] also acknowledged the importance of publication types in determining the quality of clinical evidence. They use a working definition of the 'strength of evidence' as a sum of the scores given to journal types, publication types and publication years of individual publications. Their scores are used for citation ranking, not evidence grading, and therefore their results cannot be compared to ours. However, their research does suggest that the journal names and publication years have an influence on the qualities of individual publications, which in turn may influence the grade of evidence obtained from them.

4 Methods

We used the corpus[6] proposed by Mollá [11] to collect our data. Each record in the corpus is a clinical query obtained from the 'Clinical Inquiries' section of JFP. Each query is accompanied by one or more evidence based answers and each answer is generated from one or more medical publications. Furthermore, each answer contains its SOR, a list of publication references and a brief description of the publications including their publication types. From the corpus, we collected all evidence based answers that had their SORs specified. Our final set consists of 1132 evidence based answers generated from 2713 medical documents. Of the 1132 answers, 330 are of grade A, 511 of B and 291 of C. We grouped together publication types having low frequency and similar quality levels, since it was not possible to accommodate all publication types. Our final set consisted of 11 groups of known publication types, each having a different quality level, and

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

[6] The corpus is available to the research community. The authors of this paper can be contacted for details.

1 group of unknown types, as shown in Figure 1. Based on our collected data, we considered 45.1% — the accuracy when all instances are classified as B (the majority class) — as the baseline for our experiments.

4.1 Distribution of Publication Types over SORs

In an initial analysis, we studied the distribution of publication types over the SOR grades (Figure 1). In the figure, 'Other Study' refers to low frequency studies (e.g., Observational Study), 'Other Clinical Trial' refers to clinical trials other than 'Randomised Controlled Trials' (RCT) and 'Unknown' refers to articles whose publication types are not known. A clear pattern in the distribution of publication types over SORs can be seen. For SOR A, evidence primarily comes from RCTs, Systematic Reviews and Meta-analyses, and the numbers drop significantly for other publication types. For SOR C evidence, most of the evidence comes from publications presenting expert opinion, case series/reports and consensus guidelines. The distribution for SOR B has the largest spread with Cohort studies having the highest frequency. The distributions suggest that the publication types play an important role in determining the SOR.

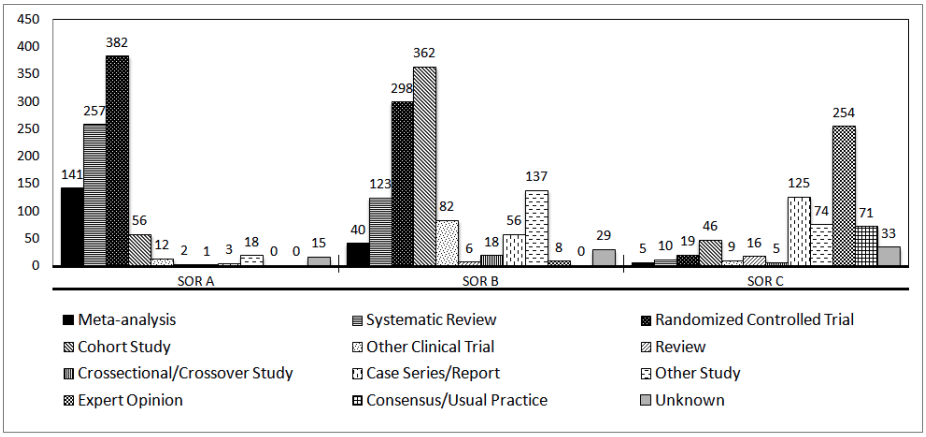

Fig. 1. Distribution of publication types across SORs.

4.2 SOR Prediction from Publication Types

To test the extent to which SORs can be predicted from the publication types, we performed basic experimentation using machine learning. We modeled the grading of evidence as a classification problem, using only the publication types of the articles as features. Each instance in our model represents an evidence based answer and is composed of the SOR class and a vector containing the

counts of each of the 12 publication types shown in Figure 1. Based on the publication types associated with each evidence, the classifiers attempt to predict the SOR (A, B or C).

We used two-thirds of our data for training and the remaining as held-out test data. For both sets, we kept the proportions of instances belonging to the three classes the same as their proportions in the whole data set. We performed our experimentation using the software package Weka[7]. Weka provides implementations of a range of classifiers organised into generic groups, and in our preliminary analysis we experimented on our training data with multiple classifiers belonging to each generic group. We chose five classifiers that produced good results on our training data and have also been shown to produce good results on similar problems in the past. The five chosen classifiers were (the names used in Weka shown in brackets): Bayes Net, SVMs (SMO), K-Nearest Neighbour (IBk), Multinomial Logistic Regression (Logistic) [10][8] and C4.5 Decision Tree (J48) [15]. For specific classifiers, we performed simple parameter tuning and chose parameter values that produced best results for stratified 10-fold cross validations on the training set. For the Bayes Net classifier, we used the K2 search algorithm [2] for local score metrics and the simple estimator for estimating conditional probability tables. For SVMs, we used John Platt's [13] sequential minimal optimisation algorithm and solved our multi-class problem using pairwise (1-vs-1) classification. We used an RBF kernel for the SVMs, normalised all attributes and used a grid search to find good values for the parameters γ and C. To find the best value of K for the K-Nearest Neighbour algorithm, we searched through all odd values of K from 1 to 101. For the C4.5 Decision Tree classifier, we searched between 2^{-5} and 2^{-1} to find the best value for the confidence factor parameter.

Classifier	**Accuracy (%)**	**95% CI**	**Parameters**
Bayes Net	66.578	61.6-71.3	$K2, SimpleEstimator$
SVMs	68.449	63.5-73.1	$\gamma = 1.0, C = 2^7$
K-Nearest Neighbour	68.717	63.8-73.4	$K = 7$
Logistic Regression	67.380	62.4-72.1	..
C4.5	68.182	63.2-72.9	$confidenceFactor = 2^{-1}$

Table 1. Accuracies, 95% confidence intervals and specific parameter values for various classifiers, using only publication types as features.

[7] http://www.cs.waikato.ac.nz/ml/weka/

[8] The Weka implementation of this algorithm is slightly different from the original implementation. Details can be found at: http://www.java2s.com/Open-Source/Java-Document/Science/weka/weka/classifiers/functions/Logistic.java.htm

4.3 SOR Prediction from other Factors

In addition to publication types, we attempted to check the influence of other factors such as journal information and publication year following Lin and Demner-Fushman's [3] work. We added the two feature sets — journal name and publication year — to our data, and performed further experimentation by adding title information of each article as a feature set. We suspected that titles may help to identify the qualities of individual publications, since they sometimes provide useful information about how the studies are carried out (e.g., 'A Double-blind, Placebo-controlled Trial'). In our model, we represented the titles and journal names using uni- and bigrams. Prior to generating the n-grams, we processed the titles by removing stop words, stemming the remaining words using the Porter stemmer and removing words occurring less than five times across the whole data set. We repeated the experimental procedures mentioned above with various combinations of these feature sets.

5 Results and Discussion

Using only publication types as a feature set, we obtained classification accuracies of approximately 66 - 69% (over 20% improvement over the baseline) with various classifiers on our held-out test set. Table 1 shows the accuracies obtained by the five above-mentioned classifiers along with 95% confidence intervals[9] for the accuracies and important parameter values for specific classifiers.

An analysis of the incorrect classifications revealed that there were few errors between A and C, which is exactly what was expected based on their very different distributions of publication types. The most common errors were between SOR A and B, and SOR C classified as B. Our manual analysis revealed that errors were caused primarily by factors such as sizes of studies, consistency and types of outcomes, which our classifiers did not take into account. For example, an essential condition for an evidence to be of grade A or B is the presence of patient-oriented outcome, irrespective of the type of study. At the same time, for certain types of publications, such as Cohort studies, the sizes of the studies significantly influence the qualities. Unaware of these information, our classifiers classified all evidences obtained primarily from Cohort studies as grade B. Furthermore, evidence obtained primarily from Meta-analyses and Systematic Reviews were graded as A, irrespective of the consistency or types of outcomes presented in the studies.

Our experiments suggest that adding factors such as journal names, publication years and article titles to the publication types do not significantly influence the SORs. Table 2 shows the highest accuracies obtained using various combinations of feature sets, the 95% confidence intervals and the classifiers producing these results. From the table it is evident that the absence of publication types as a feature set causes significant drops in accuracy. Although incorporation of article titles as a feature set produces marginally better accuracies compared to our

[9] Calculated using the package R's binom.test function (http://www.r-project.org/).

Features	Accuracy (%)	95% CI	Classifier
Journal, Pub. Year, Title and Pub. Type	63.636	58.5-68.5	C4.5
Pub. Type and Pub. Year	66.578	61.6-71.3	C4.5
Pub. Type and Title	67.380	62.4-72.1	C4.5
Pub. Type and Journal	63.904	58.8-68.8	C4.5
Journal, Pub. Year and Title	50.802	45.6-56.0	SVMs
Journal and Pub. Year	46.257	41.1-51.5	SVMs
Title only	51.070	45.9-56.2	SVMs
Pub. Year only	47.594	42.4-52.8	Bayes Net
Journal only	47.326	42.2-52.5	Bayes Net

Table 2. Accuracies, 95% confidence intervals, and best performing classifiers for various feature sets.

baseline, our experiments show that no significant improvements are achieved when this feature set is combined with publication types. The other feature sets, alone or in combination with each other, do not give a statistically significant improvement over the baseline.

6 Conclusion and Future Work

In this paper, we have discussed some experiments towards the challenging task of automatic evidence grading. Our experiments have produced encouraging results, suggesting that automatic grading of evidence is possible and modeling evidence grading as a classification problem might be an effective approach. Using publication types alone as features, it is possible to predict SORs with close to 70% accuracy. The experiments also show that information such as journal names, publication years and article titles do not significantly influence the SORs. Our manual analysis revealed that a large number of the errors are caused due to the absence of information such as study sizes and consistency among studies. Our future work will focus on incorporating these information as features. There has already been some research on polarity assessment of clinical outcomes [12], and extraction of specific information from medical abstracts (such as study sizes) [3]. We will attempt to build on these works for generating more features for our classifiers.

It would also be interesting to perform an assessment of agreement among human graders of clinical evidence. The evidence based summaries contained in JFP are prepared by domain experts, and there is a possibility that there are inconsistencies among human generated grades. Such an assessment will require significant time contribution from domain experts.

Acknowledgments

This research is jointly funded by Macquarie University and CSIRO. The authors would like to thank the anonymous reviewers for their helpful comments.

References

1. Aphinyanaphongs, Y., Tsamardinos, I., Statnikov, A., Hardin, D., Aliferis, C.F.: Text categorization models for high-quality article retrieval in internal medicine. JAMIA 12(2), 207–216 (2005)
2. Cooper, G.F., Herskovits, E.: A bayesian method for the induction of probabilistic networks from data. Machine Learning 9, 309–347 (October 1992)
3. Demner-Fushman, D., Lin, J.J.: Answering clinical questions with knowledge-based and statistical techniques. Computational Linguistics 33(1), 63–103 (2007)
4. Ebell, M.H., Siwek, J., Weiss, B.D., Woolf, S.H., Susman, J., Ewigman, B., Bowman, M.: Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. Am Fam Physician 69(3), 548–556 (Feb 2004)
5. Ely, J., Osheroff, J.A., Chambliss, M.L., Ebell, M.H., Rosenbaum, M.E.: Answering physicians' clinical questions: Obstacles and potential solutions. JAMIA 12(2), 217–224 (2005)
6. Ely, J.W., Osheroff, J.A., Ebell, M.H., Bergus, G.R., Levy, B.T., Chambliss, M.L., Evans, E.R.: Analysis of questions asked by family doctors regarding patient care. BMJ 319(7206), 358–361 (Aug 1999)
7. Goetz, T., von der Lieth, C.W.: PubFinder: a tool for improving retrieval rate of relevant PubMed abstracts. Nucleic Acids Research 33, W774–W778 (2005)
8. Greenhalgh, T.: How to read a paper: The Basics of Evidence-based Medicine. Blackwell Publishing, 3 edn. (2006)
9. Kilicoglu, H., Demner-Fushman, D., Rindflesch, T.C., Wilczynski, N.L., Haynes, B.R.: Towards automatic recognition of scientifically rigorous clinical research evidence. JAMIA 16(1), 25–31 (January 2009)
10. Le Cessie, S., Van Houwelingen, J.C.: Ridge Estimators in Logistic Regression. Applied Statistics 41(1), 191–201 (1992)
11. Mollá, D.: A Corpus for Evidence Based Medicine Summarisation. In: Proceedings of the Australasian Language Technology Association Workshop. vol. 8 (2010)
12. Niu, Y., Zhu, X., Li, J., Hirst, G.: Analysis of polarity information in medical text. In: Proceedings of the AMIA Annual Symposium. pp. 570–574 (2005)
13. Platt, J.C.: Fast training of support vector machines using sequential minimal optimization. In: Advances in Kernel Methods: Support Vector Learning. pp. 185–208. MIT Press, Cambridge, MA (1998)
14. Plikus, M., Zhang, Z., Chuong, C.M.: PubFocus: semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithm. BMC Bioinformatics 7(1), 424–439 (2006)
15. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc. (1993)
16. Sackett, D.L., Rosenberg, W.M.C., Gray, J.A.M., Haynes, R.B., Richardson, W.S.: Evidence based medicine: what it is and what it isn't. BMJ 312(7023), 71–72 (1996)
17. Sarker, A., Mollá-Aliod, D.: A Rule-based Approach for Automatic Identification of Publication Types of Medical Papers. In: Proceedings of the ADCS Annual Symposium. Melbourne, Australia (December 2010)
18. Tang, T., Hawking, D., Sankaranarayana, R., Griffiths, K., Craswell, N.: Quality-Oriented Search for Depression Portals. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) Advances in Information Retrieval, Lecture Notes in Computer Science, vol. 5478, chap. 60, pp. 637–644. Springer Berlin / Heidelberg, Berlin, Heidelberg (2009)