

Toward Automatic Recognition of High Quality Clinical Evidence

Halil Kilicoglu, MS,^{1,2} Dina Demner-Fushman, MD, PhD,² Thomas C. Rindfleisch, PhD,²
Nancy L. Wilczynski, PhD,³ R. Brian Haynes, MD, PhD³

¹Concordia University, Department of Computer Science and Software Engineering,
Montreal, Canada

²National Library of Medicine, Bethesda, MD

³Health Information Research Unit, McMaster University, Hamilton, Canada

Abstract

Automatic methods for recognizing topically relevant documents supported by high quality research can assist clinicians in practicing evidence-based medicine. We approach the challenge of identifying articles with high quality clinical evidence as a binary classification problem. Combining predictions from supervised machine learning methods and using deep semantic features, we achieve 73.5% precision and 67% recall.

Introduction

Recently, automated approaches, both knowledge-based and statistical, have shown promise in identifying high-quality articles to support evidence-based medicine. We explore supervised machine learning techniques for automatically recognizing MEDLINE[®] citations containing rigorous clinical evidence and investigate whether domain knowledge in the form of deep semantic features improves classification results.

Methods

We based our study on the test collection created to develop clinical query filters for PubMed [1]. This collection consists of 49,028 MEDLINE documents classified across three dimensions, one of which is scientific rigor (yes/no). Our training set consisted of 10,000 documents (750 rigorous), and we tested our models on 2,000 (200 rigorous) documents.

SemRep [2], a knowledge-based natural language processing system, provided the semantic features in the form of UMLS Metathesaurus concepts and Semantic Network relations between them.

We experimented with three supervised machine learning methods, naïve Bayes, polynomial support vector machine (SVM) and boosting, and an ensemble learning technique, stacking, which combined the predictions of the above three classifiers. Classification was conducted in three scenarios distinguished by features used: (1) features identified in [3] containing words from the title and abstract, MeSH indexing terms, and publication type

(baseline); (2) baseline features augmented with semantic relations; and (3) arguments of relations (UMLS concepts) added to the second feature vector. We calculated precision, recall, and their harmonic mean (F_1 score) as well as the area under the receiver operating characteristic (ROC) curve (AUC).

Results and Discussion

In the first scenario, SVM performed best in terms of AUC (0.921), while boosting yielded the best F_1 score (0.614). Stacking did not lead to any improvement. The second scenario yielded similar results on the base classifiers, while stacking outperformed all base classifiers in terms of F_1 score (0.698) and AUC (0.927). In the third scenario, the results either deteriorated or were not affected. These results indicate that deep semantic features do not significantly improve classification results over the domain knowledge encoded in MeSH terms and publication type. However, combining predictions from the base classifiers, which use these semantic features via stacking, benefits automatic recognition of high quality evidence.

References

1. Wilczynski NL, Morgan D, Haynes RB; Hedges Team. An overview of the design and methods for retrieving high-quality studies for clinical care. *BMC Med Inform Decis Mak.* 2005; 5:20.
2. Rindfleisch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypemymic propositions in biomedical text. *J Biomedical Informatics.* 2003; 36(6): 462-77.
3. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text Categorization Models for High-Quality Article Retrieval in Internal Medicine. *J Am Med Inform Assoc.* 2005; 2: 207-16.