UIC/OHSU CLEF 2018 Task 2 Diagnostic Test Accuracy Ranking using Publication Type Cluster Similarity Measures

Aaron M. Cohen[1] and Neil R. Smalheiser[2]

[1]Oregon Health & Science University, Portland, Oregon, USA
[2] University of Illinois College of Medicine, Chicago, Illinois, USA
cohenaa@ohsu.edu, neils@uic.edu

Abstract. The CLEF 2018 Task 2 goal was to identify and rank retrieved articles relevant to conducting a systematic diagnostic test accuracy review on a given topic. The UIC/OHSU team did not attempt to rank retrieved articles by relevance directly, but rather explored the baseline value of ranking retrieved articles according to the probability that they are concerned with diagnostic test accuracy.

First, a set of six publication type clusters, including a cluster of diagnostic test accuracy papers (DTAs), was built by searching PubMed from 1987-2015. We created several types of cluster similarity measures for each publication type. Similarity types included: implicit-term similarity, most important word similarity, journal similarity, and author count similarity. These similarity features were then used with weighted and un-weighted linear SVM machine learning algorithms, which were trained with a data set retrieved from PubMed searches consisting of 3481 PMIDS likely to be DTAs, and 71684 PMIDS most of which are not likely to be DTAs. The trained models produce scores predicting the probability that an individual article is a DTA. The CLEF 2018 Task 2 Test PMIDs for each topic were scored and ranked, and the cutoff probability for each of the two models determined by visual inspection of the score distribution on the test data. Cutoff probabilities chosen were 0.20 for the unweighted SVM model and 0.40 for the weighted SVM model.

Keywords: Machine Learning, Support Vector Machine, Publication Types, Diagnostic Test Accuracy.

1 Introduction

We participated in Task 2 of the CLEF 2018 e-Health challenge [1][2]. The goal of this task was to identify and rank articles relevant to conducting a systematic diagnostic test accuracy review on a given topic, among those articles returned by topic-specific PubMed queries.

Search	Query
#8	Search #7 AND #5 NOT #6
#7	Search "diagnostic test accuracy"[ti] OR "diagnostic accuracy"[ti]
#6	Search editorial[pt] OR letter[pt] OR "practice guideline"[pt] OR review[pt]
#5	Search #1 AND #2 AND #3 AND #4
#4	Search humans[MeSH Terms]
#3	Search hasabstract
#2	Search ""english""[Language]) OR ""english abstract""[Publication Type]"
#1	Search ""1987/1/1""[Date - Publication] : ""2015/12/31""[Date - Publication]

Figure 1. PubMed query used to retrieve likely DTAs for training data.

We have been extending our prior work on probability based tagging for specific publication types [3] by developing a general system to predict probabilities for multiple publication types simultaneously [4]. We applied a preliminary version of that system on six clinical publication types, reporting here only on DTA publications.

2 Methods

The UIC/OHSU CLEF 2018 Task 2 submission applies a machine learning approach to ranking the PMIDs retrieved by CLEF for 20 topics. The approach assigns probabilities to individual PMIDs based the likelihood that they are DTAs. To generate positive training data, likely DTAs were retrieved using the PubMed query shown in Figure 1. No specific information about the topic queries generating the PMID list for each query was used.

The system builds a predictive model in stages. First, publication type clusters, including diagnostic test accuracy papers (DTAs), were built by searching PubMed from 1987-2015. Six publication type (PT) clusters were used in this model: DTAs, Randomized Controlled Trials, Cross-sectional Studies, Cross-over Studies, Cohort Studies, and Case-Control Studies. These clusters were used as training data to create several types of cluster similarity measures for each publication type. The PT clusters are treated as consensus profiles that represent the PT as a whole, so any given article is judged to belong to it if it is sufficiently similar in its weighted sum of similarity features. While the members of each cluster are very likely to be examples of the cluster specific publication type, nothing in the method requires all the articles in a cluster to be of that publication type. Somewhat noisy training data is expected.

Similarity types used as features included: implicit term similarity, most important word similarity, journal similarity, and author count similarity. Implicit term similarity measures how similar a paper is to a cluster based on terms (words, bigrams, etc.) that commonly occur with words contained in the papers within each cluster relative to the baseline frequency across MEDLINE. A cluster "centroid-like" vector is computed as the mean vector of the individual cluster article vectors, where each article vector consists of the 300 weighted terms most associated to the words in the article. The cluster centroid is limited to the 300 highest total scoring terms across the cluster. See [5] for a complete and detailed description.

Most important word similarity measures the fraction of words in the paper that are in the list of most important words computed for each cluster, as measured by the frequency of the word occurring in that cluster versus MEDLINE as a whole.[6] Journal similarity measures how representative an article's journal is for a cluster, again as measured by the frequency of the journal occurring in that cluster versus the rest of MEDLINE. A MeSH based journal distance measure was used for papers published in journals that did not occur in the cluster to estimate cluster similarity based on the most similar journal in the cluster [7]. The author count similarity measures how selective the author count of a paper is for a particular cluster. Note that the criteria used for defining DTAs by PubMed search were NOT directly used by the features used in the classification model. Individual publication MeSH terms were not used directly as features in any of the similarity measures.

The four similarity measures produce one feature for each of the six publication type clusters, resulting in 24 similarity-based features. These similarity features were then used with weighted and un-weighted linear SVM machine learning algorithms, which were trained with a data set retrieved from the 1987-2015 PubMed searches. The DTA cluster was used as positive training data set, and the other clusters were combined into the negative training data set. This resulted in training data consisting of 3481 PMIDs likely to be diagnostic test accuracy papers (DTAs), and 71684 PMIDs most of which are not likely to be DTAs.

The trained weighted and un-weighted SVM models were then applied to the CLEF 2018 task 2 challenge data. The PMIDs supplied in the topic files were used to retrieve the full PubMed XML record for these articles, and the XML records used to compute the 24 similarity features for input to the trained models.

The trained models produce probability scores predicting whether or not an individual PMID is a DTA. The PMID predictions were then organized according to the CLEF 2018 Task 2 topics, and were ranked within a topic by probability. The cut-off probability for each of the two models was determined by visual inspection of the score distribution on the test data. Cutoff probabilities chosen were 0.20 for the unweighted SVM model and 0.40 for the weighted SVM model. This information was combined into the submission qrel files, rank ordering the topic publication PMIDs highest to lowest predicted probability, one file for each model. In this manner we produced two sets of predictions, submitted as two separate runs: OHSU_UIC_LIBLINW for the weighted model, and OHSU_UIC_LIBLINB for the unweighted model.

3 Results

The official overall evaluation results for our systems are shown in Table 1. Across the board, the liblinear system with inverse class frequency weighting performed slightly better than the liblinear with bias version. These results are averages across all the topics. Based on the similar CLEF 2017 task, these results are about median as compared to other entries. The average precision achieved by the our liblinear weighted system was 0.180, which would have ranked 14th out of 33 CLEF 2017 entries [8].

4 Discussion

Considering that we only ranked articles according to their probability of being a DTA, and did not evaluate query topic information at all, our approach did have some significant value in identifying articles that are relevant for inclusion in topic-specific systematic reviews.

We plan on continuing to work on our system, expanding the number of clusters and publication types, as well as add additional cluster similarity measures. While the current approach uses an SVM in a one-versus-rest approach for multi-classification, we are also experimenting with other classifiers which are more flexible with multiple category classification such as random forests and deep learning neural networks.

Table 1. Official evaluation overall results for the UIC/OHSU Task 2 system entries.

Run Label	**OHSU_UIC_LIBLINW**	**OHSU_UIC_LIBLINB**
Algorithm	Liblinear with inverse frequency class weights	Liblinear with bias term
WSS@100%	0.164	0.154
WSS@95%	0.264	0.255
Recall@10%	0.296	0.289
Recall@20%	0.473	0.462
Recall@30%	0.579	0.562
Recall@40%	0.641	0.624
Recall@50%	0.695	0.683
Recall@60%	0.751	0.739
Recall@70%	0.805	0.793

Recall@80%	0.860	0.846
Recall@90%	0.935	0.926
Recall@100%	1.000	1.000
Average Precision	0.180	0.174

References

1. Suominen H, Kelly L, Goeuriot L, Kanoulas E, Azzopardi L, Spijker R, et al. Overview of the CLEF eHealth Evaluation Lab 2018. In: CLEF 2018 - 8th Conference and Labs of the Evaluation Forum. CEUR-WS: Springer; 2018. (Lecture Notes in Computer Science (LNCS)).
2. Kanoulas E, Spijker R, Li D, Azzopardi L. CLEF 2018 Technology Assisted Reviews in Empirical Medicine Overview. In: CLEF 2018 Evaluation Labs and Workshop. CEUR-WS; 2018.
3. Cohen AM, Smalheiser NR, McDonagh MS, Yu C, Adams CE, Davis JM, et al. Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine. J Am Med Inform Assoc JAMIA. 2015 May;22(3):707–17.
4. Smalheiser NR, Cohen AM. Design of a generic, open platform for machine learning-assisted indexing and clustering of articles in PubMed, a biomedical bibliographic database. Data Inf Manag. 2018;2(1):1–10.
5. Smalheiser NR, Bonifield G. Unsupervised Low-Dimensional Vector Representations for Words, Phrases and Text that are Transparent, Scalable, and produce Similarity Metrics that are Complementary to Neural Embeddings. ArXiv Prepr ArXiv180101884. 2018;
6. Smalheiser NR, Zhou W, Torvik VI. Distribution of "Characteristic" Terms in MEDLINE Literatures. Information. 2011;2(2):266–76.
7. Jennifer LD, Smalheiser NR. Three journal similarity metrics and their application to biomedical journals. PloS One. 2014;9(12):e115681.
8. Kanoulas E, Li D, Azzopardi L, Spijker R. CLEF 2017 technologically assisted reviews in empirical medicine overview. In: CEUR Workshop Proceedings. 2017. p. 1–29.