Semi-Supervised Information Retrieval System for Clinical Decision Support

Harsha Gurulingappa[1]*, Alexander Bauer[1§], Luca Toldo[1]*
Claudia Schepers* and Gerard Megaro[+]

*Merck KGaA, Frankfurterstraße 250, 64297 Darmstadt, Germany
[§]Cognizant Technology Solutions, Torhaus Westhafen, Speicherstraße 57-59, 60327 Frankfurt am Main, Germany
[+]EMD Millipore Corporation, 290 Concord Road, MA-01821 Billerica, USA

Abstract

This article summarizes the approach developed for TREC 2016 Clinical Decision Support Track. In order to address the daunting challenge of retrieval of biomedical articles for answering clinical questions, an information retrieval methodology was developed that combines pseudo-relevance feedback, semantic query expansion and document similarity measures based on unsupervised word embeddings. The individual relevance metrics were combined through a supervised learning-to-rank model based on gradient boosting to maximize the normalized discounted cumulative gain (nDCG). Experimental results show that document distance measures derived from unsupervised word embeddings contribute to significant ranking improvements when combined with traditional document retrieval approaches.

1. Introduction

The goal of the TREC 2016 Clinical Decision Support Track[2] is to retrieve biomedical articles relevant for answering clinical questions based on patient records. The target document collection is a snapshot of the Open Access Subset of PubMed Central (PMC) from March 28, 2016, containing 1,251,954 full-text articles. Participants were tasked to retrieve articles useful for answering questions related to three types of generic clinical questions:

- Diagnosis: What is the patient's diagnosis?
- Test: What tests should the patient receive?
- Treatment: How should the patient be treated?

Three versions of the patient records were provided. The note format represents the content of the history of present illness (HPI) section of the electronic health record (EHR). The description format provides a simplified and shortened version of the HPI, removing abbreviations and jargon. The summary format is a

[1] Authors contributed equally. Contact Email: jerry.megaro@emdmillipore.com
[2] http://www.trec-cds.org/2016.html

condensed representation of the description, giving a 1-2 sentence summary of the case. For each type of clinical question, 10 topics were provided containing summary, description and notes summing to overall 30 topics. Submissions of retrieved and ranked articles were judged by medical librarians and physicians and retrieved documents were classified as "definitely relevant", "potentially relevant" or "definitely not relevant".

In this report, described are the methodologies used for document indexing (Section 2.1), query expansion (Section 2.2), pseudo relevance feedback (section 2.3), supervised ranking based on unsupervised word embeddings (section 2.4 and 2.5), and results (section 3).

2. Methods

2.1 Document Indexing and Retrieval

The document collection for 2016 consists of 1,251,954 PMC articles in NXML format that were indexed with Solr (version 5.5.2). In order to facilitate supervised learning component of the workflow, document collections from 2014 and 2015 tracks were separately indexed. Document collections for 2014 and 2015 were identical with 733,138 PMC articles published until January 21, 2014. Fields used for indexing were PMCID, title, abstract, body, conclusion, journal title and journal type. Indexing was performed with standard Solr settings which include tokenization, stemming and stopword removal. A master stopword list was constructed which is a combination of standard English stopwords and a list of stopwords constructed manually based on most frequent terms occurring in the document collection. Query parser applied was Extended DisMax[3] (eDisMax) which is an improved version of DisMax query parser with improved features such as advanced stopword handling and improved proximity boosting. Okapi BM25[4] was the similarity measure used for querying and retrieval from the index. For each of the indices having document collections from 2014, 2015 and 2016, identical parameters were applied for indexing, stopword handling, query parsing and retrieval scoring.

2.2 Semantic Query Expansion with UMLS

Observations from previous TREC-CDS tracks indicated that query expansion with UMLS [Bodenreider2004] concepts can significantly contribute to improved retrieval results [Palotti2015]. Therefore, the MetaMap program [Aronson2010] was applied for the identification of UMLS concepts in topics. Since UMLS has over 100 semantic types[5], mappings were restricted to only the following semantic types:

> *Disease or Syndrome, Sign or Symptom, Pathologic Function, Diagnostic Procedure, Anatomical Abnormality, Laboratory Procedure, Pharmacologic Substance, Neoplastic Process, Therapeutic or Preventive Procedure*

[3] https://cwiki.apache.org/confluence/display/solr/The+Extended+DisMax+Query+Parser
[4] http://nlp.stanford.edu/IR-book/html/htmledition/okapi-bm25-a-non-binary-model-1.html
[5] https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html

Selection of semantic types was based on manual curation of results of MetaMap applied on topics from 2014 and 2015 tracks. For all the mapped concepts, synonyms were extracted from UMLS knowledge sources[6] which were thereafter used for query expansion. Query terms from UMLS were weighted lower than the topic terms with a weight of *0.1* for each term.

2.3 Pseudo-Relevance Feedback

For a given query, Pseudo-Relevance Feedback[7] (PRF) was implemented to collect words from titles of top *k* retrieved documents. Stopwords were eliminated and the remaining words were added to the initial query in order to generate a new query that can be reused for searching and retrieval. Additional query terms generated as a result of PRF were weighted lower than topic terms with a weight of *0.1* for each term. Experiments were conducted by varying values of *k* on the 2014 and 2015 datasets, and the optimal parameters were chosen.

2.4 Document Distances from Unsupervised Word Embeddings

Commonly used document distance metrics for document retrieval are typically based on a bag-of-words (BoW) approach or a term-weighting approach (like TF-IDF, BM25). One important drawback of these methods is that distances between words are not taken into account, thereby misrepresenting the distance between documents that carry the same information but use different words. In order to circumvent the drawbacks associated with classical BoW approaches, word embeddings were incorporated as an additional method in conjugation with supervised learning-to-rank (*see Section 2.5*). Latent low-dimensional document representations like Latent Semantic Indexing (LSI), Latent Dirichlet Allocation (LDA) do not suffer from earlier mentioned limitation and particularly recently developed word embedding methods have shown to deliver better accuracy on distance-based tasks [Kusner2015].
To leverage on these methods for the medical document retrieval task, cosine distances between document centroids of topics and articles were computed for use in a downstream supervised learning-to-rank model. Centroids for articles were calculated based on the abstract, the title or the journal title. Word vectors were obtained from both publicly available embeddings (*i.e.* Wikipedia) as well as corpus derived from 2016 document collection. All word embeddings were generated using the GloVe model [Pennington2014]. The corpus was tokenized and lowercased, no stemming was performed. The models were trained for 100 iterations using a window size of 10 and vector dimension of 300.

[6] https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html
[7] http://nlp.stanford.edu/IR-book/html/htmledition/pseudo-relevance-feedback-1.html

2.5 Supervised Learning-to-Rank Model

Supervised learning-to-rank models compute optimized ranking functions based on training data that has been reviewed and annotated for relevance by human assessors. They provide an effective tool to optimally combine unsupervised ranking functions and have been successfully applied in previous year's clinical decision support track [Song2015]. Additionally, they enable design of topic-independent features that incorporate article meta-data such as publication type and design of features that account for the type of medical question.

For learning the ranking function, gradient boosting was applied to maximize normalized discounted cumulative gain (nDCG) using LambdaRank gradient approximation [Burges2006] which is implemented as part of the machine learning library XgBoost [Chen2016]. The library supports both linear models as well as decision trees as weak learners. Topics and document relevance scores from 2014 and 2015 served as training and validation data for feature selection and hyper-parameter optimization.

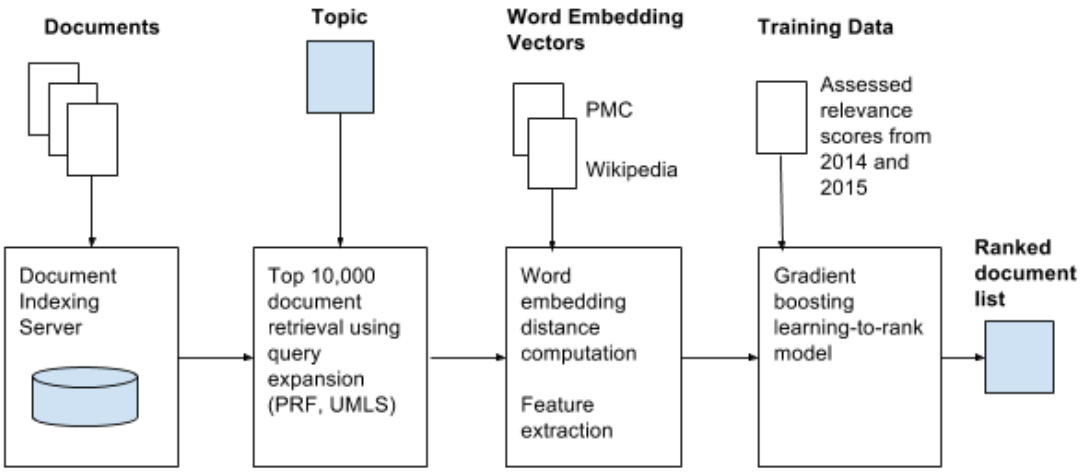

Figure 1: Overview of workflow implemented

3. Results

3.1 Performance Evaluation

Performances of retrieved documents were evaluated using standard TREC retrieval measures i.e. inferred normalized Discounted Cumulative Gain (infNDCG) as a primary metric and inferred Average Precision (infAP) as a secondary metric [Clough2013]. Topics and indices of document collections from 2014 and 2015 tracks formed the training set. Topics from 2014 and 2015 were separately used for cross validation of results on the training data. Various experiments were conducted with different retrieval strategies and respective parameter settings. Their influence on results were studies and the best possible settings were applied on 2016 topics and dataset.

3.2 Pseudo-Relevance Feedback

Performance of PRF depends on top *k* retrieved documents. For cross validation, performance was measured using various values of *k* i.e. 10, 20, 30, 40 and 50. Although, the value of *k* had only minor impact on retrieval performance, *k=30* was chosen which was observed to deliver the best results. For the runs that leveraged on UMLS query expansion, terms in the topic after stopword removal in combination with UMLS terms (i.e. expanded queries) were used as initial query.

3.3 Learning-to-rank model optimization

A gradient boosting learning-to-rank technique was applied to re-rank the initially retrieved set of 10000 documents per topic. Different base learners and regularization methods were explored for gradient boosting. Performance of boosting with linear base model and L1 regularization was on-par with a regression tree base. To reduce the risk of overfitting the training set, a linear base model was chosen for the final submission. The best cross-validation score was achieved from 100 iterations at a learning rate of 0.1 and L1 regularization alpha=0.001.

Feature engineering made a significant difference in the performance. Ranking scores obtained from PRF, UMLS expansion and word embedding document distances were used as features. In addition, features were added to account for relevance bias on ‘article type’ and occurrence of keywords related to the medical question of the topic (i.e. ‘diagnosis’, ‘treatment’, ‘test’). Adding feature interactions turned out to be beneficial when using the linear model, particularly multiplication of BM25 scores and word embedding distances. Ultimately, feature sets described in Table 1 contributed to increased performance in iterative experiments and were included in the final model. BM25 and word embedding distances of the pre-selected 10,000 documents per topic were normalized to zero mean and unit variance.

Feature set F.1:

F.1.1 BM25 from PRF query expansion

F.1.2 BM25 from PRF query expansion combined with UMLS query expansion

Feature set F.2:

Word embedding document distances between:

F.2.1 topic and article title using PubMed vectors

F.2.2 topic and journal title using PubMed vectors

F2.3 topic and article title using Wikipedia vectors

Feature set F.3:

F.3.1 Product of feature F.1.1 and F.2.1

Feature set F.4:

F.4.1-F4.29 Dummy binary variables for article type (29 types)

Feature set F.5:

Binary variables indicating whether title contains at least one question related word:

F.5.1 Question is 'treatment' and the title contains 'treatment', 'efficacy', 'therapy', 'management' or 'surgery'

F.5.2 Question is 'diagnosis' and the title contains 'diagnos', 'symptom', 'detect', 'identif', 'characteristic', 'laboratory', 'parameters' or 'predict'

F.5.3 Question is 'test' and title contains 'test'

Table 1: Feature sets used for learning-to-rank

3.4 Performance of Runs

Table 2 describes the topic fields and methodologies used for various runs submitted in 2016. Table 2 provides a comparison of infNDCG scores for runs using topics from 2014 to 2016. For the runs submitted in 2016, Solr search with topic terms and PRF was used as baseline (MRKSumCln). Additional runs were submitted with UMLS query expansion on top of PRF (MRKUmlsSolr) and with learning-to-rank model (MRKUmlsXgb). Although the methodology was not optimized for searching with notes, MRKPrfNote run was submitted using terms in notes as input for PRF, however excluding learning-to-rank since no prior training data was available.

Run	Topic Field	PRF Query Expansion	UMLS Query Expansion	Learning-to-Rank Model
MRKPrfNote	Notes	X		
MRKSumCln	Summary	X		
MRKUmlsSolr	Summary	X	X	
MRKUmlsXgb	Summary	X	X	X

Table 2: Description of runs submitted to TREC CDS 2016

Run	2014	2015 (A)	2016
MRKPrfNote	-	-	0.1504
MRKSumCln	0.2229	0.2672	0.2223
MRKUmlsSolr	0.2321	0.2724	0.2261
MRKUmlsXgb	0.2368	0.2769	0.2493
Median	*0.151*	*0.2288*	*0.1858*

Table 3: Evaluation of retrieval results (infNDCG scores) for all TREC CDS topics from 2014 to 2016

Observations from Table 3 and Table 4 shows that query expansion with UMLS contributed consistently to the retrieval performance. Learning-to-rank method boosted the performance of runs significantly especially in 2016. One reason could be because of the double the size of training data compared to 2014 and 2015 cross validation. Overall, the baseline as well as best performing runs were better than the median of performances of all submissions.

Run	Topic Field	infNDCG	infAP
MRKPrfNote	Notes	0.1504	0.0179
MRKSumCln	Summary	0.2223	0.0272
MRKUmlsSolr	Summary	0.2261	0.0285
MRKUmlsXgb	Summary	0.2493	0.0315

Table 4: Comparison of infNDCG and infAP scores for different runs for 2016

3.5 Manual Assessment of Selected Run

Topic-20	Summary: A 87 yo female reports several days abdominal pain, worse yesterday, severe and more localized to the right, accompanied by nausea and vomiting. Labs show elevated bilirubin, transaminitis, amylase and lipase.								
	PMCID	Female	Abdominal Pain	Nausea	Vomiting	Elevated Bilirubin	Elevated Transaminitis	Elevated Amylase	Elevated Lipase
Hit-1	4332755	X	X	X	X	X	X	X	X
Hit-2	2923793	X	X	-	X	X	-	X	X
Hit-3	4275784	X	X	X	X	X	X	X	X
Hit-4	2779365	-	X	X	X	-	-	X	X
Hit-5	3878513	X	X	X	X	X	X	-	-

Table 5: Manual assessment of retrieval results for the run MRKUmlsXgb Topic-20

A knowledge-based manual assessment was performed for retrieval results of one of the selected runs (MRKUmlsXgb Topic:20) which achieved highest infNDCG of 0.7 (*See Table 5*). Top 5 hits were manually evaluated for relevancy by two biomedical experts. Results of assessment showed that top five ranked documents were relevant to search criteria indicative of patients with abdominal pain undergoing laboratory tests and showing variations in blood protein levels.

4. Conclusion

A methodology for information retrieval from full-text scientific articles for answering clinical questions has been presented. Semantic query expansion showed consistent improvement of results in comparison to classical keyword-based retrieval. A supervised learning-to-rank technique based on unsupervised word embedding features significantly contributed to the performance of retrieval. Best observed results with formerly described methodology measured in terms of infNDCG is significantly higher than median infNDCG of results submitted by all participants for 2016.

Although, state-of-the-art components have been applied such as BM25 for retrieval scoring, word embeddings for computing special similarity of terms and learning-to-rank for document ranking optimization, several parameters can be further investigated to check their impact on retrieval performance. For instance, BM25 was used as retrieval scoring function since queries were executed against entire document. Multi-Field search with BM25F scoring function can be evaluated by treating and prioritizing different fields in documents such as title, abstract and body separately. Current retrieval approach does not account for semantic relationships or sentence level proximity of keywords. An experiment indexing and retrieval over semantic relationships or sentence level co-occurrence statistics is worth for investigation.

Similarly, various other components used in the workflow related to word embeddings and feature sets used for learning-to-rank including algorithm parameters can be fine-tuned to understand their impact on performance.

Overall, the presented approach has indicated good results and outperformed the current track's median with infNDCG higher than 6%. It has exhibited competence to change the way biomedical information can be retrieved with higher accuracy. This can enable biomedical and clinical researchers to fetch most relevant information for their needs thereby substantially reducing the overall time and effort needed for manually filtering the irrelevant information or repeated searching to ensure maximum recall.

The developed methodology can have several real world uses case scenario. For instance, the system could be trained and applied in clinical recommendation systems for suggesting most relevant medical records for a given patient case optimized for physician needs. Due to the domain independence and easy adaptability of the developed approach, it can be further applied for information retrieval from various other document sources such as patents for enabling prior art search and technology scouting.

5. References

[Goodwin2014] Goodwin T. and Harabagiu S. M. UTD at TREC 2014: Query expansion for clinical decision support. In Proc. TREC 2014, 2014. Online: http://trec.nist.gov/pubs/trec23/papers/pro-UTDHLTRI_clinical.pdf

[Kusner2015] Kusner M. J. and Sun Y. and Kolkin N. I. and Weinberger, K. Q., From Word Embeddings To Document Distances, In Proc. ICML 2015, Online: http://jmlr.org/proceedings/papers/v37/kusnerb15.pdf

[Mikolov2013] Mikolov T., Yih W., and Zweig G. 2013. Linguistic Regularities in Continuous Space Word Representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2013), pages 746–751, Atlanta, US.

[Pennington2014] Pennington J., Socher R., and Manning C. 2014. GloVe: Global vectors for word representation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), pages 1532–1543, Doha, QA.

[Song2015] Song Y., He Y., Hu Q., and He L. (2015). ECNU at 2015 CDS Track: Two Re-ranking Methods in Medical Information Retrieval. In Proceedings of the 2015 Text Retrieval Conference. Online: http://trec.nist.gov/pubs/trec24/papers/ECNU-CL.pdf

[Burges2006] Burges C. J., Ragno R., and Le Q. V. "Learning to rank with nonsmooth cost functions," in NIPS 2006, pp. 395–402, 2006. Online: http://research.microsoft.com/en-us/um/people/cburges/papers/LambdaRank.pdf

[Chen2016] Chen T. and Guestrin C. (2016). "XGBoost: A Scalable Tree Boosting System". In: Proceedings of the ACM SIGKKD International Conference on Knowledge Discovery and Data Mining. KDD '16.

[Palotti2015] Palotti J. and Hanbury A. "TUW@TREC Clinical Decision Support Track 2015". Preceedings of TREC Conference. 2015.

[Bodenreider2004] Bodenreider O. "The Unified Medical Language System (UMLS): integrating biomedical terminology". Nucleic Acids Res. 2004 Jan 1; 32(Database issue): D267–D270.

[Aronson2010] Aronson A.R., Lang F.M. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010 May-Jun;17(3):229-36.

[Clough2013] Clough P. and Sanderson M. Evaluating the performance of information retrieval systems using test collections. Information Research. vol. 18 no. 2, June, 2013.