Comparing Word Embeddings for Document Screening based on Active Learning

Andres Carvallo and Denis Parra[0000−0001−9878−8761]

Computer Science Department
Pontificia Universidad Catolica de Chile
Santiago, Chile
afcarvallo@uc.cl, dparra@ing.puc.cl

Abstract. Document screening is a fundamental task within Evidence-based Medicine (EBM), a practice that provides scientific evidence to support medical decisions. Several approaches are attempting to reduce the workload of physicians who need to screen and label hundreds or thousands of documents in order to answer specific clinical questions. Previous works have attempted to semi-automate document screening, reporting promising results, but their evaluation is conducted using small datasets, which hinders generalization. Moreover, some recent works have used recently introduced neural language models, but no previous work have compared, for this task, the performance of different language models based on neural word embeddings, which have reported good results in the latest years for several NLP tasks. In this work, we evaluate the performance of two popular neural word embeddings (Word2vec and GloVe) in an active learning-based setting for document screening in EBM, with the goal of reducing the number of documents that physicians need to label in order to answer clinical questions. We evaluate these methods in a small public dataset (HealthCLEF 2017) as well as a larger one (Epistemonikos). Our experiments indicate that Word2vec have less variance and better general performance than GloVe when using active learning strategies based on uncertainty sampling.

Keywords: active learning · evidence based medicine · document screening · word embeddings

1 Introduction

Evidence-based Medicine (EBM) is a practice that provides scientific evidence to support medical decisions. This evidence nowadays is obtained from biomedical journals, usually accessible through the portal PubMed[1], a search engine which provides free access to abstracts of biomedical research articles as well as to the MEDLINE database. An existing problem is to find relevant documents within a massive volume of documents, given a clinical question or a query. As a consequence of this, the time required for the search and screening of articles

[1] https://www.ncbi.nlm.nih.gov/pubmed/

related to clinical questions about medical problems can take long and sometimes it consumes a large part of a physician's workday [15, 6]. When people conduct this repetitive task, there is a good chance of overlooking important articles, which can have a negative impact on decisions such as the patient's treatment [12]. Moreover, the publication of medical papers has grown exponentially the last decade. Since 2005, PubMed has indexed more than 1 million articles per year, which means that the process of searching and manual screening of medical evidence will become increasingly more difficult for physicians without the support of information retrieval and machine learning algorithms. For this reason, some systems have emerged to support experts in the collection of evidence such as Embase[2], DARE[3] and Epistemonikos[4]. In this article, we work with data from Epistemonikos, which helps expert physicians to review and validate scientific evidence grouped by medical questions to facilitate its subsequent search. Our goal is to improve the efficiency and efficacy of document screening in the practice of EBM. In other words, we aim at reducing the effort made by physicians at screening documents to find the evidence needed to support the answers of a medical question. We use an active learning approach, experimenting with a large dataset of medical questions, unlike previous works which use very small datasets, some of them very recent [13]. In this short paper, we contribute by: *i*) Experimenting in both a large dataset (Epistemonikos, 987) and a small dataset (CLEF, 50), showing evidence of generalization of our approaches, and *ii*) comparing the performance of documents represented with two state-of-the-art neural word embeddings (Word2vec [14] and GloVe [18]) as well as traditional relevance feedback [5].

2 Related Work

The task of finding relevant documents related to a medical question through citation screening has been studied and it is known as the *total recall problem*: given a medical topic or question, find all the documents that are relevant about a particular topic. Recently, the CLEF task 2 [11] is a challenge that calls for solving the problem of prioritizing which documents to screen to reduce work overload for experts. They provide a public dataset with medical topics and a set of candidate documents; participants have to rank documents by relevance for every specific medical subject in the minimum of iterations to make more efficient the document screening process. In the literature, the approaches to solving this problem are based on two general lines: information retrieval and machine learning methods.

In the **information retrieval** area, there have been many attempts to solve the problem using techniques such as relevance feedback [5], query expansion [13], ranking and inference based on external knowledge [8]. However, they do not

[2] https://www.elsevier.com/solutions/embase-biomedical-research
[3] https://www.crd.york.ac.uk/CRDWeb/
[4] https://www.epistemonikos.org/en

ensure a level of recall necessary to capture all the evidence related to a medical question.

From the **machine learning** community, the approach is to automate or semi-automate the screening process or review of medical articles that were previously selected as relevant to a medical question by learning the pattern of physicians conducting a document survey. There have been efforts to solve this problem by using automatic classification [2, 3, 1, 16, 21]. Where they compared classifiers such as Naive Bayes, K-NN, and SVM, using different ways to represent text, such as word embeddings and bag-of-clinical terms from titles and abstracts.

There is also literature that has used active learning [9, 7, 22, 15] for medical topic detection and clinical text classification. Moreover, a few of deep learning models have been proposed for the classification of relevant evidence and categorization of documents in medical questions [4, 10]. Generally, the majority of work done has used datasets of up to 50 medical topics/questions and 200,000 documents, and in this case, we work with a dataset close to $1,000$ medical questions and $370,000$ potential documents, allowing models to generalize and obtain better efficacy results compared to the state of the art. In addition, unlike previous work, we compare two neural word embedding models for document representation (Word2vec [14] and GloVe [18]) in order to asses their performance for the biomedical document screening task.

3 Proposed Solution

The process of finding documents that answer a clinical question requires first retrieving a set of candidate documents. Then, physicians perform the document screening where they verify that abstracts and titles of each document are related to the medical question, this particular process may involve much time and cognitive effort from experts.

Problem formulation: Given a medical question q and a set of candidate documents $C = \{c_1, c_2, ...c_n\}$ we need to ask an oracle (physician) O to label these documents as relevant or not relevant to q. We want to avoid asking the labeling of every document, so we select an informative sample to be labeled by the expert. With these labels, we train a predictive model M. It might be necessary to ask for labels in many iterations in order to refine the model, ending up with several models $M_0, M_1, ..., M_k$.

In our case, we use an active learning (AL) approach [20]. Using an AL strategy A (e.g., uncertainty sampling, query-by-committee, etc.), we sample a set of unlabeled documents X from C, in order to ask the oracle O to label them. With the labeled items, we then train a machine learning model $M_i(X, Y)$ with the new observations $X \subset C$ with labels Y (binary, $Y = 1$ means relevant document and $Y = 0$ means not relevant) given by O. Then, we use the trained model M_i to predict relevance labels for unobserved documents, and using the active learning strategy A we select new items to be labeled by O in order to create an updated version of the model M_{i+1}. In each iteration, we evaluate the

model (precision, recall,), and we can stop until a fixed number of iterations or after the model converges.
To address the problem above, we developed a system, where we start with a small proportion of labeled documents as relevant or not relevant for each medical question to train a first version of the machine learning model M_i. Then, using the active learning strategy we chose instances to be labeled by a physician based on the title and abstract text features represented internally as word embeddings (GloVe and Word2vec). After the physician adds the labels, they are used to train a machine learning model M_{i+1} to predict the relevance of new unlabeled documents and thus begin a new iteration.
The performance of our approach first depends on the machine learning algorithm chosen and second, on the active learning strategy that chooses unlabeled examples to create a labeled dataset as input for supervised learning algorithms. The strategies used in this experiment are uncertainty sampling and random sampling, given their lower complexity compared to others such as error-based, gradient-based and variable reduction [20]. The machine learning algorithms that are considered to be trained with new labeled examples are random forests, logistic regression, and neural networks.
With respect to the active learning sampling strategies, *random sampling:* chooses random documents to train the machine learning model and it is usually used as a comparison baseline against other approaches. On the other side, *uncertainty sampling:* looks for records which have higher label prediction uncertainty, making them potentially more informative for collecting their actual labels and then training or updating a model.

4 Experiments

Dataset. For the experiments, we used two datasets: CLEF[5] and Epistemonikos. Both of them have a similar distribution of documents per question, where the majority of medical questions contain an approximate number of 200 relevant documents. On the one hand, CLEF dataset contains only 50 medical questions and 200,000 documents related to them that were crawled from PubMed using each document id. On the other hand, the Epistemonikos Evidence Synthesis Project is a collaborative initiative established in 2012 with the objective of collecting, organizing and comparing all relevant evidence for health decision-making, through a multilingual platform. This dataset is composed of 987 medical questions and 372,829 potential documents. In both datasets each medical question is associated to a Systematic Review (hereinafter, SR), which is a type of article that collects and synthesizes the most relevant primary studies and trials related to a question. The information of documents from both datasets consists of the title, abstract, author, year and the label if it is relevant (or not) to the question or medical subject. In the case of Epistemonikos data, the labels were previously curated by senior medical students, in which they had to select papers related to a set of medical questions. *Document representation*: for each

[5] https://sites.google.com/site/clefehealth2017/task-2

document we lower case the concatenation of title and abstract, then remove stop words, and we use GloVe [18] and Word2vec [14] to obtain an embedding representation of 300 dimensions of each word. Finally, using average pooling we obtain a vector for each document.

4.1 Offline Active Learning Setup

We experimented doing a simulation of the active learning labeling process of documents for medical questions. As each medical question has a different number of relevant documents, we sample documents that are not relevant where the total of relevants corresponds to the 5%, so that the distribution of documents is similar to the CLEF dataset [13]. We filtered out some of the medical questions, keeping those that have more than five relevant documents and less than 2,000 relevant documents, ending up with 987. We compared the results of applying active learning on the CLEF dataset that contains 50 SR (Systematic Reviews) with Epistemonikos. For each medical question, we hide the document labels and we leave only five with their respective labels to start building the model and then iterating with active learning to receive feedback from the oracle. For each prediction made by the machine learning model in each iteration, we sorted the results depending on the predicted probability of being relevant for each model, so the evaluation metrics were calculated with the ranked list of potential candidates given by each strategy. The parameters chosen for machine learning algorithms were: for neural networks we used five hidden layers, ReLu activation function, learning rate of 1e-05, momentum of 0.9, 100 neurons per layer and Adam optimization function. For the random forest, we used 100 estimators. Experiments were programmed in Python3 using libact [23], scikit-learn [17], pandas and gensim libraries. Code for these experiments will be published in a github repository after notification.
Evaluation metrics. We evaluated our proposed active learning strategies with traditional IR metrics also used by Lee et al. [13]: precision@k, recall@k and mean average precision (MAP). We report the metrics obtained after ten iterations, with ten documents labeled per iteration.

5 Results and Discussion

Table 1 presents the results. The first column indicates the dataset as well as the type of embeddings. The second column shows the active learning strategy, as well as the learning model. Then the following seven columns show recall at three cut off levels (R@10, R@20, R@30), precision at three cut off levels (P@10, P@20, P@30), and Mean average precision (MAP). As shown in Table 1, for the Epistemonikos dataset, uncertainty sampling based on RF is a clear winner for recall@10 which means that this strategy captures relevant documents in the first ten positions for both GloVe and Word2vec representations of titles and abstracts. In the case of the HealthCLEF dataset, the model that achieves the best results on recall@10 is also RF, followed by NN. If we compare the performance of both word embeddings, we observe that in general, Word2vec has

Table 1: Average results with standard deviation, of recall@k (R@k), precision@k (Pr@k) and Mean Average Precision (MAP) performance measured in Epistemonikos and CLEF datasets using active learning strategies (US: uncertainty sampling, RS: random sampling) using a batch of 10 documents per feedback iteration for Word2vec and GloVe representation.

Dataset	AL-Model	R@10	R@20	R@30	Pr@10	Pr@20	Pr@30	MAP
Epistemonikos	US-NN	0.377 (0.02)	0.542 (0.04)	0.627 (0.05)	0.856 (0.045)	0.747 (0.053)	0.654 (0.053)	0.900 (0.001)
987 SRs	US-RF	0.414 (0.03)	0.590 (0.06)	0.679 (0.08)	0.926 (0.058)	0.799 (0.075)	0.696 (0.078)	**0.975** (0.001)
GloVe 300 dim	US-LR	0.307 (0.03)	0.435 (0.04)	0.498 (0.05)	0.760 (0.047)	0.670 (0.057)	0.587 (0.058)	0.875 (0.001)
	RS-LR	0.052 (0.001)	0.094 (0.003)	0.127 (0.004)	0.413 (0.012)	0.362 (0.01)	0.315 (0.01)	0.625 (0.07)
Epistemonikos	US-NN	0.391 (0.02)	0.562 (0.04)	0.645 (0.05)	0.877 (0.04)	0.760 (0.05)	0.663 (0.05)	0.932 (0.02)
987 SRs	US-RF	**0.417** (0.03)	**0.596** (0.05)	**0.687** (0.07)	**0.934** (0.04)	**0.807** (0.06)	**0.704** (0.06)	0.973 (0.001)
Word2vec 300 dim	US-LR	0.413 (0.02)	0.593 (0.03)	0.684 (0.04)	0.912 (0.04)	0.791 (0.04)	0.691 (0.04)	0.958 (0.02)
	RS-LR	0.021 (0.01)	0.054 (0.002)	0.125 (0.002)	0.283 (0.01)	0.192 (0.01)	0.293 (0.01)	0.463 (0.05)
CLEF	US-NN	0.427 (0.01)	0.573 (0.01)	0.665 (0.01)	0.841 (0.02)	0.782 (0.02)	0.702 (0.01)	0.935 (0.02)
50 SRs	US-RF	0.416 (0.01)	0.583 (0.01)	0.688 (0.01)	0.865 (0.01)	0.871 (0.01)	**0.729** (0.01)	0.965 (0.01)
GloVe 300 dim	US-LR	0.146 (0.019)	0.206 (0.04)	0.228 (0.03)	0.758 (0.014)	0.555 (0.04)	0.446 (0.07)	0.957 (0.04)
	RS-LR	0.033 (0.01)	0.077 (0.002)	0.099 (0.002)	0.392 (0.01)	0.278 (0.01)	0.322 (0.01)	0.374 (0.05)
CLEF	US-NN	0.402 (0.01)	0.552 (0.01)	0.687 (0.01)	0.865 (0.01)	0.753 (0.01)	0.726 (0.01)	0.985 (0.01)
50 SRs	US-RF	**0.428** (0.01)	**0.586** (0.01)	**0.689** (0.01)	0.867 (0.01)	0.785 (0.01)	0.713 (0.01)	**0.989** (0.01)
Word2vec 300 dim	US-LR	0.170 (0.026)	0.249 (0.047)	0.297 (0.06)	**0.892** (0.018)	**0.805** (0.018)	0.723 (0.017)	0.930 (0.08)
	RS-LR	0.019 (0.01)	0.045 (0.002)	0.095 (0.002)	0.192 (0.01)	0.382 (0.01)	0.273 (0.01)	0.172 (0.05)
Epistemonikos	Rel. Feed. (Rocchio)	0.28	0.35	0.42	-	-	-	0.45
TF-IDF	BM25	0.18	0.25	0.32	-	-	-	0.35

better and more stable performance. GloVe present more substantial variations in different ML models.

6 Conclusion and Future Work

In this article we supported results from previous studies in terms of showing that active learning with an uncertainty sampling strategy yields good results for the task of biomedical document screening. Moreover, we contribute by comparing two popular word embeddings to represent documents: Word2vec and GloVe. The best results were obtained using Word2vec document representation and random forests as the learning algorithm. GloVe document representation also yields competitive results, but it seems more sensitive two the classification model used: it performs well with random forests but shows poor performance with neural networks and logistic regression. Moreover, our experiments indicate that these results are consistent in both the small public dataset of HealthCLEF and the larger dataset of Epistemonikos, giving evidence of generalization.
For future work, we will try other machine learning models, active learning strategies and evaluate the results using CLEF metrics [11]. We will also test other paradigms for more scalable learning, such as weak supervision. With respect, to embeddings, we will test different values of sensitive parameters, as mentioned by Roy et al. [19]. Finally, we will conduct a user study with actual physicians in order to evaluate online the performance of our approach.

7 Acknowledgements

We acknowledge Epistemonikos Foundation, the Chilean research agency Conicyt, Fondecyt grant 1191791 and the Millenium Institute IMFD.

References

1. Adeva, J.G., Atxa, J.P., Carrillo, M.U., Zengotitabengoa, E.A.: Automatic text classification to support systematic reviews in medicine. Expert Systems with Applications **41**(4), 1498–1508 (2014)
2. Bekhuis, T., Tseytlin, E., Mitchell, K.J., Demner-Fushman, D.: Feature engineering and a proposed decision-support system for systematic reviewers of medical evidence. PloS one **9**(1), e86277 (2014)
3. Choi, S., Ryu, B., Yoo, S., Choi, J.: Combining relevancy and methodological quality into a single ranking for evidence-based medicine. Information Sciences **214**, 76–90 (2012)
4. Del Fiol, G., Michelson, M., Iorio, A., Cotoi, C., Haynes, R.B.: A deep learning method to automatically identify reports of scientifically rigorous clinical research from the biomedical literature: Comparative analytic study. J Med Internet Res **20**(6) (Jun 2018)
5. Donoso-Guzmán, I., Parra, D.: An interactive relevance feedback interface for evidence-based health care. In: 23rd International Conference on Intelligent User Interfaces. pp. 103–114. ACM (2018)
6. Elliott, J.H., Turner, T., Clavisi, O., Thomas, J., Higgins, J.P., Mavergames, C., Gruen, R.L.: Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. PLoS medicine **11**(2), e1001603 (2014)
7. Figueroa, R.L., Zeng-Treitler, Q., Ngo, L.H., Goryachev, S., Wiechmann, E.P.: Active learning for clinical text classification: is it better than random sampling? Journal of the American Medical Informatics Association **19**(5), 809–816 (2012)
8. Goodwin, T.R., Harabagiu, S.M.: Knowledge representations and inference techniques for medical question answering. ACM Transactions on Intelligent Systems and Technology (TIST) **9**(2), 14 (2018)
9. Hashimoto, K., Kontonatsios, G., Miwa, M., Ananiadou, S.: Topic detection using paragraph vectors to support active learning in systematic reviews. Journal of biomedical informatics **62**, 59–65 (2016)
10. Hughes, M., Li, I., Kotoulas, S., Suzumura, T.: Medical text classification using convolutional neural networks. Stud Health Technol Inform **235**, 246–50 (2017)
11. Kanoulas, E., Li, D., Azzopardi, L., Spijker, R.: Clef 2017 technologically assisted reviews in empirical medicine overview. In: CEUR Workshop Proceedings. vol. 1866, pp. 1–29 (2017)
12. Keselman, A., Smith, C.A.: A classification of errors in lay comprehension of medical documents. Journal of biomedical informatics **45**(6), 1151–1163 (2012)
13. Lee, G.E., Sun, A.: Seed-driven document ranking for systematic reviews in evidence-based medicine. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 455–464. ACM (2018)
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
15. Miwa, M., Thomas, J., OMara-Eves, A., Ananiadou, S.: Reducing systematic review workload through certainty-based screening. Journal of biomedical informatics **51**, 242–253 (2014)
16. Mo, Y., Kontonatsios, G., Ananiadou, S.: Supporting systematic reviews using lda-based document representations. Systematic reviews **4**(1), 172 (2015)

17. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of machine learning research **12**(Oct), 2825–2830 (2011)
18. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
19. Roy, D., Ganguly, D., Bhatia, S., Bedathur, S., Mitra, M.: Using word embeddings for information retrieval: How collection and term normalization choices affect performance. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. pp. 1835–1838. ACM (2018)
20. Settles, B.: Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning **6**(1), 1–114 (2012)
21. Wallace, B.C., Small, K., Brodley, C.E., Lau, J., Schmid, C.H., Bertram, L., Lill, C.M., Cohen, J.T., Trikalinos, T.A.: Toward modernizing the systematic review pipeline in genetics: efficient updating via data mining. Genetics in medicine **14**(7), 663 (2012)
22. Wallace, B.C., Small, K., Brodley, C.E., Trikalinos, T.A.: Active learning for biomedical citation screening. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 173–182. ACM (2010)
23. Yang, Y.Y., Lee, S.C., Chung, Y.A., Wu, T.E., Chen, S.A., Lin, H.T.: libact: Pool-based active learning in python (2017)