Using Deep Learning to Identify Critical Documents for Clinical Decision Support Systems

Brandon Fan[1], Long Xia[2], Patrick Weiguo Fan[2, 3], Edward A. Fox[4]

[1]Blacksburg High School

[2]Center for Business Intelligence and Analytics, CBIA

[3]Accounting and Information Systems, ACIS

[4]Department of Computer Science, CS

Virginia Polytechnic Institute and State University

Keywords*:* Clinical Decision Making, Deep Learning, Learn2Rank

Abstract

Doctors/Physicians are challenged with effective clinical decision making regarding the treatment plans for patients with specific conditions/symptoms. They often resort to clinical decision support systems to help them come up with the best treatment plan for patients at critical times. However, the search quality of current clinical decision support systems is often low, so they fail to help doctors find relevant medical articles related to their patients' conditions. To help improve search ranking performance in clinical decision support systems, we introduce a novel deep-learning (DL) based learning-to-rank algorithm that can retrieve more relevant and important biomedical articles matching a doctor's search queries containing patients' conditions or symptoms. We compared the performance of the DL-based algorithm to multiple benchmarks (including state-of-the-art system implementations for this task) and found that we achieve better results. The newly designed ranking algorithm can be incorporated into existing clinical decision support systems to assist doctors in making better and more informed clinical decisions, reduce medical costs, and ultimately save patients' lives.

1. Introduction & Problem

Because of problems with current search engines and algorithms, physicians and doctors need a more efficient and reliable method to search biomedical articles to identify the optimal treatment for each of their clinical patients. A Clinical Decision Support System (CDS) can be defined as a system that assists doctors in determining the most beneficial treatment for a patient with particular conditions. This system includes receiving a query from a doctor, retrieving relevant documents, and then sorting or ranking the documents in accordance to their pertinence or relevance. However, despite the copious amounts of biomedical articles and data, current CDS and medical search engines are far from optimal. Oftentimes, these systems return non-relevant articles to doctors. One reason for this is due to a poor ranking algorithm (an algorithm that estimates the relevance of a particular article to a specific query). One recent study in 2016 of a CDS by Zhang et al. achieved only 3% in precision (a metric of relevance with maximum of 100%).

Therefore, our goal is to use a deep learning-based algorithm to find an efficient and effective ranking algorithm for biomedical articles to improve clinical decision support systems.

The rest of this paper analyzes our solution to this problem. First, we discuss the significance of our research, current algorithms, and our strategy to solve this problem. Then we examine the data and the technical details involved in solving this problem, our performance, and comparisons with other algorithms. Lastly, we share our insights gained and plans for future work.

2. Research Significance

This problem of ranking motivates us to find a better algorithm. One method to solve this problem is with the use of deep learning. Deep learning has gained traction in many fields such as image classification (He et al., 2015), image captioning (Yang et al., 2017), and more. Due to the success in multiple disciplines, deep learning provides a great prospect for a ranking function necessary to return and sort relevant results to physicians and doctors. Without a fine-tuned ranking algorithm, a clinical decision support system will be unable to determine which documents are relevant. Thus, the ranking algorithm plays a pivotal role in the entire clinical decision support system as it estimates the relevance of a document.

Prior research in ranking algorithms for CDS has not fared well. Precision scores (how many retrieved documents are actually relevant out of the documents returned) are very low with standard IR models, reaching around only 3% in one CDS study (Zhang et al., 2016). The algorithms considered simply analyze text from a lexical and syntactical standpoint and do not take into account semantic relationships between words in an article. Though some prior research has used semantic relationships in conjunction with standard algorithms through the use of embeddings, precision scores still are only 26%, indicating most of the articles returned are not relevant to the doctor's query (Jo, Lee, 2016). The utilization of deep learning techniques, including convolutional neural networks (CNN) and convolutional long short-term memory networks (C-LSTM) for classification, with fine-tuned word representations specifically for the medical corpus (Mikolov et al., 2013), and other deep-learning based algorithms, have not been considered for this task. Therefore, we introduce a novel solution that utilizes deep learning in ranking in clinical decision support systems.

We:

1. **Utilize richer semantic word relationships through the fine-tuning of word vectors using a medical corpus.**
2. **Investigate the potentiality of a deep learning-based ranking algorithm and how results compare with prior research.**

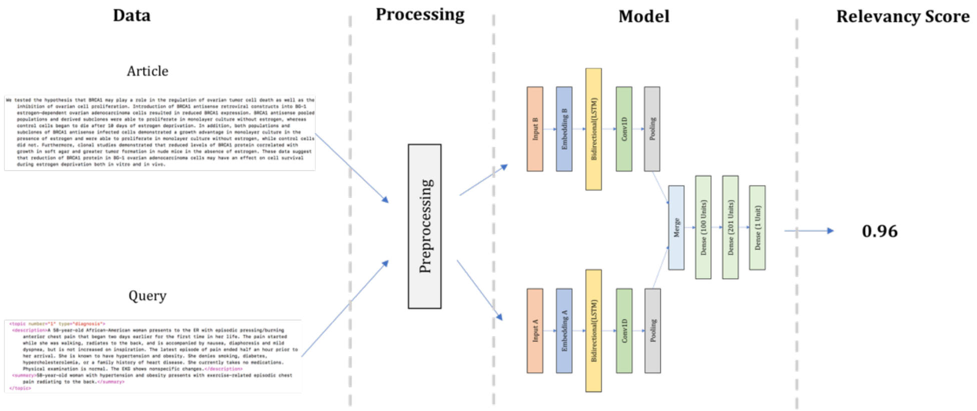
This task is very challenging, as prior research has shown. One reason is due to the variation in article lengths. Another challenge is the use of medical acronyms and jargon that is unrecognizable unless one is working in the field. Because of the nature of such text, it is a challenge to design an algorithm that can recognize and handle these variations during processing.

3. Related Work

Various algorithms have been proposed to solve the ranking problem in CDS. The ranking algorithm must be able to distinguish between relevant and non-relevant documents, quickly and correctly. Many standard algorithms in the information retrieval field have been utilized. One is the use of the term frequency, inverse document frequency (TF-IDF) vectorization ranking algorithm. Variations of the algorithm include MATF and Okapi Best Matching 25 (BM25) (Zhang et al., 2016). Pre-trained word embeddings (vector representations of words) including the Global Vectors (GLoVe) word embeddings (Manning et al., 2013) were used to create the basis of a semantic text vectorization algorithm to compose vectors of the PubMed (Public Medical Database) and TREC corpus' and construct a cosine similarity matrix for text relevancy (Jo, Lee, 2016) to then be utilized by standard ranking algorithms like BM25. Other ranking algorithms follow a hierarchical design. A hierarchical design of document clustering (termed MeSH or Medical Subject Headings). MeSH is a classification system designed to use hierarchical clustering of different disease levels (i.e. endocrine system is a level 1 cluster, thyroid disease is a level 2 cluster, and thyroid dysgenesis as a final bottom level cluster) (Jo, Lee, 2016). A final method is the use of a ranking support vector machine (SVM) classification algorithm to classify documents into various subsets (Li et al., 2016).

4. Research Approach

We will use deep learning algorithms to attempt to solve the clinical decision support ranking problem. In this section, we introduce the dataset and preprocessing methods, the deep learning model architecture, and the validation metrics utilized for the learning to rank task. The approach is shown below in **Figure 1**.



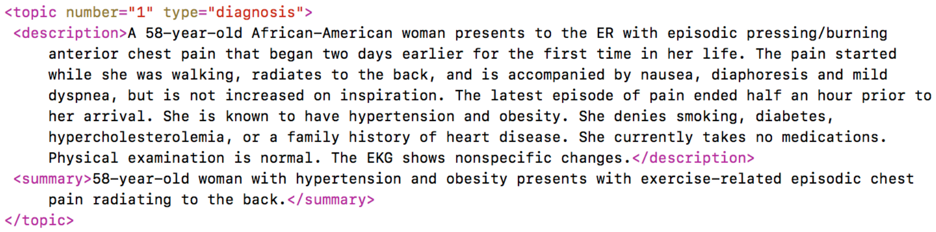
We tested the hypothesis that BRCA1 may play a role in the regulation of ovarian tumor cell death as well as the inhibition of ovarian cell proliferation. Introduction of BRCA1 antisense retroviral constructs into BG-1 estrogen-dependent ovarian adenocarcinoma cells resulted in reduced BRCA1 expression. BRCA1 antisense pooled populations and derived subclones were able to proliferate in monolayer culture without estrogen, whereas control cells began to die after 10 days of estrogen deprivation. In addition, both populations and subclones of BRCA1 antisense infected cells demonstrated a growth advantage in monolayer culture in the presence of estrogen and were able to proliferate in monolayer culture without estrogen, while control cells did not. Furthermore, clonal studies demonstrated that reduced levels of BRCA1 protein correlated with growth in soft agar and greater tumor formation in nude mice in the absence of estrogen. These data suggest that reduction of BRCA1 protein in BG-1 ovarian adenocarcinoma cells may have an effect on cell survival during estrogen deprivation both in vitro and in vivo.

```
<topic number="1" type="diagnosis">
 <description>A 58-year-old African-American woman presents to the ER with episodic pressing/burning
    anterior chest pain that began two days earlier for the first time in her life. The pain started
    while she was walking, radiates to the back, and is accompanied by nausea, diaphoresis and mild
    dyspnea, but is not increased on inspiration. The latest episode of pain ended half an hour prior to
    her arrival. She is known to have hypertension and obesity. She denies smoking, diabetes,
    hypercholesterolemia, or a family history of heart disease. She currently takes no medications.
    Physical examination is normal. The EKG shows nonspecific changes.</description>
 <summary>58-year-old woman with hypertension and obesity presents with exercise-related episodic chest
    pain radiating to the back.</summary>
</topic>
```

Figure 1: Research Workflow

4.1 TREC Dataset & Preprocessing

In order to train the algorithm towards the task of clinical decision support, we utilized the text retrieval conference (TREC) clinical decision support track dataset. TREC is an annual hosted competition for various applications or "tracks" of text retrieval. The clinical decision support track ran from 2014 to 2016. Each year, the data is structured around query statements from doctors and physicians. The task contains 30 queries from doctors and an associated 10,000 documents per query, each with a binary label (0 or 1) of relevancy (is it relevant or not). This results in a total of 90 queries over 3 years and 900,000 documents in total. A query is shown below in eXtensible Markup Language (XML) as well as a distribution of text lengths in the dataset.



```
<topic number="1" type="diagnosis">
 <description>A 58-year-old African-American woman presents to the ER with episodic pressing/burning
     anterior chest pain that began two days earlier for the first time in her life. The pain started
     while she was walking, radiates to the back, and is accompanied by nausea, diaphoresis and mild
     dyspnea, but is not increased on inspiration. The latest episode of pain ended half an hour prior to
     her arrival. She is known to have hypertension and obesity. She denies smoking, diabetes,
     hypercholesterolemia, or a family history of heart disease. She currently takes no medications.
     Physical examination is normal. The EKG shows nonspecific changes.</description>
 <summary>58-year-old woman with hypertension and obesity presents with exercise-related episodic chest
     pain radiating to the back.</summary>
</topic>
```

Figure 2: Query in XML

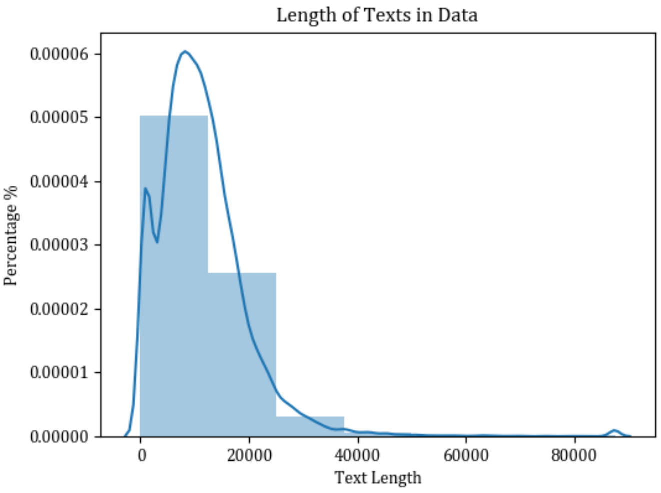


Figure 3: Text Length Distribution

Due to the high skewness of the text lengths and the vast amounts of jargon in the articles, multiple methods of preprocessing must be utilized to handle the text length and acronyms. To preprocess the biomedical articles and queries, we utilized an English stop words list to remove all stop words, lower-cased all characters, and then lemmatized all words to their root form (ex., Swimming to swim). After, the data is tokenized (split into an array of words and then split into training, validation, and testing sets for the deep learning model to learn, validate, and test upon.

4.2 Converting Words to Vectors

In order for the deep learning model to be able to learn the representations for the medical articles, the articles and words from the preprocessing step (4.1) must be mapped into a vector space. This is done through the use of word embeddings. We utilized the GloVe and fine-tuned medical embeddings for experimentation and performance. Utilizing an embedding will convert each word into a point in a vector space so that the deep learning model can computationally act upon the text corpus.

4.3 A Deep Learning Architecture for Ranking

Our deep learning model consists of common layers found present in the deep learning field and industry. The model receives two input vectors, a query vector and a biomedical article vector. The model then moves these through a word embedding layer which performs the actions described above (4.1). The results of the embedding layer are fed into a bidirectional Long-Short Term Memory (LSTM) layer, and then into a convolution and a max pooling layer. The results are then merged and fed into multiple fully connected layers, and a final sigmoid neuron calculates a relevance score between 0 and 1. The loss function to be used during gradient descent is the binary cross entropy loss function defined by the equation below.

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(h_\theta(x_i)\right) + (1 - y_i)\log\left(1 - h_\theta(x_i)\right)$$

Equation 1: Binary Cross Entropy Loss Function

The entire architecture from a visual perspective is shown below:

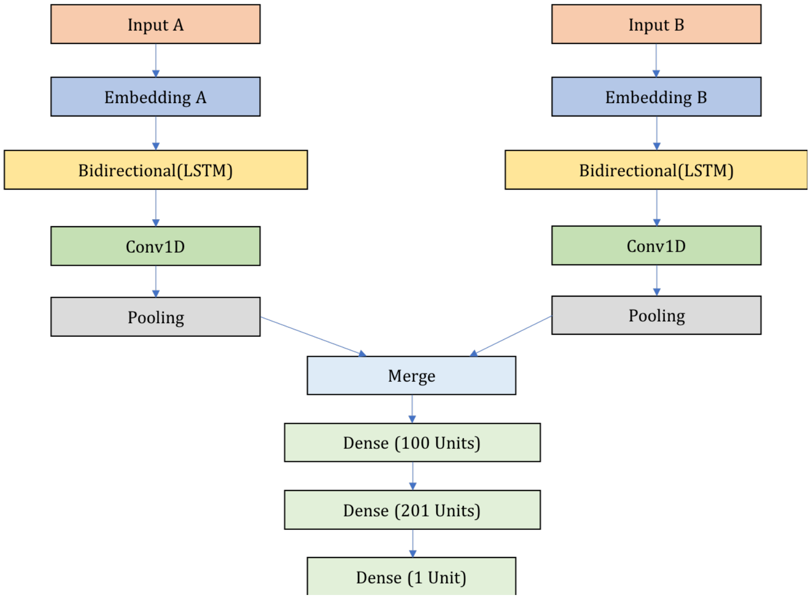


Figure 4: Model Architecture

The model is trained on 1/3 of the queries from each year (2014, 2015, and 2016), validated on 1/3 of the queries each year, and tested on 1/3 of the queries each year. The model is trained for 60 epochs. Dropout was used as the primary regularization method between Convolutions and Dense layers.

4.4 Evaluation Metrics

Four metrics were utilized to determine model efficacy and accuracy:

precision for 1000 documents (P@1000), normalized discounted cumulative gain for 1000 documents (NDCG@1000), average precision (AP, precision for multiple document sizes), and the receiver operating characteristic (ROC).

Precision can be stated as the metric that measures "the percentage of relevant documents returned by the model that are actually relevant". The higher the precision, the better the model. Normalized discounted cumulative gain measures the relevancy score of search results and grades the resulting documents for "is the proper percentage given to a highly relevant document? And, are the documents returned actually relevant?". This metric combines the precision with the idea that better search results should have higher relevancy scores. The higher the better. Average precision is the precision calculated at various document counts (e.g., P@10, P@100, etc.) and averaged to determine an overall precision. The higher the better. Finally, ROC measures the ability for a model to discriminate between relevant and non-relevant documents at various thresholds; it results in a metric between 0 and 1, the higher the better.

5. Experimental Results

Two deep learning models were trained, one with the GloVe word embeddings trained on Wikipedia articles and one with custom fine-tuned medical embeddings trained the TREC queries and documents. The fine-tuned medical embeddings model (Our Model in the table) results are compared with benchmark methods in the table and chart below.

Model Name	**NDCG@1000**	**Average Precision**	**P@1000**	**ROC**
LR-CV*	0.592888	0.633281	0.608	0.516679
LR-TFIDF*	0.574253	0.650065	0.609	0.484234
SVM-W2V*	0.49838	0.639181	0.495	0.516771
LR-W2V*	0.664953	0.665832	0.671	0.536941
Our Model	**0.783463**	**0.675371**	**0.768**	**0.680792**

**LR: Logistic Regression, CV: Count Vectorization, TF-IDF: Term Frequency-Inverse Document Frequency, SVM: Support Vector Machine, W2V: Word2Vec*

Table 1: Evaluation of Table Results

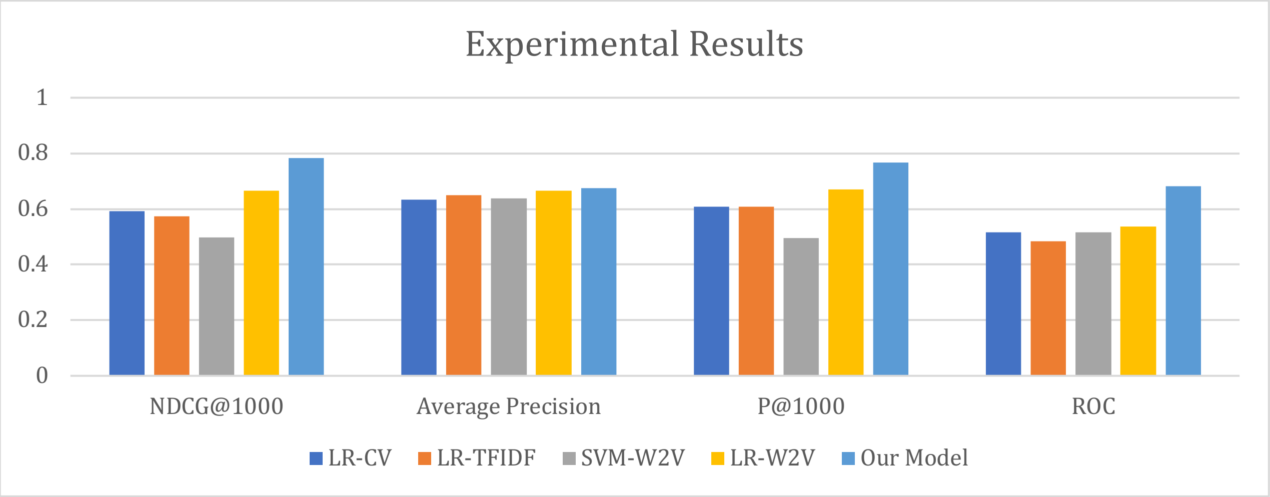


Figure 5: Experimental Results Chart

As shown in **Table 2** and **Figure 5**, the deep learning model performs better than current information retrieval methods. The results show an increase in performance with the use of a deep learning-based model. The results beat prior research in all metrics. This shows that a deep learning model is a viable and a better option for ranking and returning documents relevant to doctors and physicians.

6. Conclusion & Further Work

Though results were positive, much more work can be done to improve the deep learning model. Due to the imbalanced data, better text processing and feature engineering can be implemented so that the deep learning model is fed useful and feature-rich data. Utilization algorithms that can synthetically generate new samples of text data will be of significant help to increasing the sample size for the model to learn upon (Chawla et al., 2011) which can improve performance. The deep learning model can also be trained to generalize better to unseen datasets. Future work should also investigate the use of shallower networks (less layers) as well as a custom loss function that is specific to the task of CDS ranking. Using pre-trained weights trained upon other sources of data such as the standard Web-100K information retrieval dataset may increase the performance of the ranking algorithm. Finally, the use of an ensemble-based model that utilizes our deep learning model combined with other state-of-the-art information retrieval algorithms to compute a relevancy score may result in a more efficient and more accurate model for the task of estimating relevance relative to a query.

7. References

Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyr, W. P. (2011). *SMOTE: Synthetic Minority Over-sampling Technique.* Retrieved from https://arxiv.org/abs/1106.1813

He K., Zhang X., Ren S., Sun J. (2015). *Deep Residual Learning for Image Recognition*. Retrieved from https://arxiv.org/abs/1512.03385

Karimi, S., Falamaki, S., Nguyen V. (2016). *CSIRO at TREC Clinical Decision Support Track*. Retrieved from http://trec.nist.gov/pubs/trec25/papers/CSIROmed-CL.pdf

Li B., Xiao J., Jiang S., Wang Z., Ding H., Niu Y. (2016). *Literature based Clinical Decision Support: Searching Articles According to Medical Needs*. Retrieved from http://trec.nist.gov/pubs/trec25/papers/HAUT-CL.pdf

Jo S., Lee K. (2016). *CBNU at TREC 2016 Clinical Decision Support Track.* Retrieved from http://trec.nist.gov/pubs/trec25/papers/cbnu-CL.pdf

Zhang Y., Jian F., Hu P. (2016). *CCNU at TREC 2016 Clinical Decision Support Track*. Retrieved from http://trec.nist.gov/pubs/trec25/papers/CCNU2016TREC-CL.pdf

Mikolov T., Chen K., Corrado G., Dean J. (2013). *Efficient Estimation of Word Representations in Vector Space*. Retrieved from https://arxiv.org/abs/1301.3781

Pennington J., Socher R., Manning C. D. (2013). *GloVe: Global Vectors for Word Representation*. Retrieved from https://nlp.stanford.edu/pubs/glove.pdf

Xie J., Girshick R., Farhadi A. (2016). *Unsupervised Deep Embedding for Clustering Analysis*. Retrieved from https://arxiv.org/pdf/1511.06335.pdf

Yang Z., Zhang Y., Rehman S., Huang Y. (2017). *Image Captioning with Object Detection and Localization*. Retrieved from https://arxiv.org/pdf/1706.02430.pdf