electronics

MDPI

*Article*

# Automatic Identification of High Impact Relevant Articles to Support Clinical Decision Making Using Attention-Based Deep Learning

**Beomjoo Park [1,†], Muhammad Afzal [2,†], Jamil Hussain [1], Asim Abbas [1] and Sungyoung Lee [1,*]**

[1] Department of Computer Science & Engineering, Kyung Hee University, Yongin 446-701, Korea; pbj@oslab.khu.ac.kr (B.P.); jamil@oslab.khu.ac.kr (J.H.); asimabbasturi@oslab.khu.ac.kr (A.A.)

[2] Department of Software, Sejong University, Seoul 05006, Korea; mafzal@sejong.ac.kr

[*] Correspondence: sylee@oslab.khu.ac.kr; Tel.: +82-31-201-2514

[†] These authors contributed equally.

Received: 5 July 2020; Accepted: 20 August 2020; Published: 22 August 2020

**Abstract:** To support evidence-based precision medicine and clinical decision-making, we need to identify accurate, appropriate, and clinically relevant studies from voluminous biomedical literature. To address the issue of accurate identification of high impact relevant articles, we propose a novel approach of attention-based deep learning for finding and ranking relevant studies against a topic of interest. For learning the proposed model, we collect data consisting of 240,324 clinical articles from the 2018 Precision Medicine track in Text REtrieval Conference (TREC) to identify and rank relevant documents matched with the user query. We built a BERT (Bidirectional Encoder Representations from Transformers) based classification model to classify high and low impact articles. We contextualized word embedding to create vectors of the documents, and user queries combined with genetic information to find contextual similarity for determining the relevancy score to rank the articles. We compare our proposed model results with existing approaches and obtain a higher accuracy of 95.44% as compared to 94.57% (the next best performer) and get a higher precision by about 14% at P@5 (precision at 5) and about 12% at P@10 (precision at 10). The contextually viable and competitive outcomes of the proposed model confirm the suitability of our proposed model for use in domains like evidence-based precision medicine.

**Keywords:** machine learning; deep learning; precision medicine; clinical decision support; healthcare; health management; health communication

## 1. Introduction

Plenty of research has been dedicated to designing solutions to obtain a better score in the identification of high impact studies in PubMed literature [1–3] and on the topic of matching query-document pairs to encourage the document ranking results [3,4]. The recent development in modern medicine compelled medical professionals to look for relevant information in the secondary databases. However, doubts regarding the inaccurate results of searching and retrieval prevent them from adapting using such solutions in their usual clinical practices. Notably, in precision medicine, the automatic identification of expressed genes and the studies in which they are reported to become more critical for tailored diagnosis, prognosis, and treatment strategies. The search for medical documents on a patient's disease or condition has been around for quite some time. However, there are two problems with traditional medical document retrieval methods compared to document retrieval required by precision medicine. The first is the correct identification of disease one is looking for, and the second is the exact identification of demographics of the patient associated with a specific gene. The accuracy of finding medical documents that are appropriate for the individual needs of precision medicine may be reduced without classifying them based on disease.

This study aims to identify the relevant medical documents related to the disease by classifying documents into health condition groups, followed by searching for gene and patient demographics. More specifically, the objectives of this work are; (a) explore the effectiveness of attention-based deep learning models for the classification and feature vector creation tasks in the biomedical domain and (b) investigate whether additional considerations of query matching with documents would affect search results and performance. We propose a Bidirectional Encoder Representation from Transformers (BERT) based model, train on a dataset consisting of 240,324 clinical articles, which is collected from 2018 Precision Medicine track in Text REtrieval Conference (TREC) to identify and rank relevant documents that matched the query. First, prior to searching through a query, a pre-classification was performed, the data was divided into five health condition classes. For classification, we use two textual features: title and abstract. Second, we used vectorized document titles and abstracts using contextualized word embedding and then match user queries with text using contextual similarity. Third, we employed the Okapi best matching algorithm called BM25 to calculate the importance of genetic information in each document that corresponds to the topics of interest provided by TREC. Finally, we calculated the final ranking score by summing the contextual similarity score and BM25 score.

In summary, the main research contributions of this work are: (a) design of a sequential two-stage approach of identifying relevant articles in a large corpus of biomedical articles, (b) implementation of a self-learning-based deep learning model using BERT pre-trained classification, (c) automatic generation of queries containing gene and demographic information, and (d) semantic matching of query and text using contextual word embeddings for relevant document ranking.

## 2. Background and State of the Art

This work is conducted to identify relevant articles for a given set of topics in the data provided by TREC 2018 Precision Medicine Track. TREC, co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense, was started in 1992 to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies [5]. Every year it organizes a workshop consisting of a set of tracks, areas of focus in which particular retrieval tasks are defined. Our work considers the precision medicine track offered in 2018, which focuses on a critical use case in clinical decision support: providing useful precision medicine-related information to clinicians treating cancer patients. It uses synthetic examples created by precision oncologists at the University of Texas MD Anderson Cancer Center. Each case describes the patient's disease (e.g., a type of cancer), the relevant genetic variants (which genes), and basic demographic information (age, sex). The cases are semi-structured and require minimal natural language processing. The challenge of this track is to retrieve biomedical documents in the form of article abstracts to address relevant treatment information for the given patient.

### 2.1. TREC Evaluation

The TREC evaluation named "trec_eval" is a standard tool used to evaluate the ranking of documents based on relevance [6]. The assessment is based on two files; the first file, known as "qrels" (query relevance), lists the relevant decisions for each query. The second is a "result" file containing the ranking of the documents returned from the information retrieval (IR) system. The "qrels" file is the correct answer file provided by TREC that contains a list of documents that are considered relevant for each query. This file can be regarded as the "correct answer" as the human makes it, and the documents retrieved by the IR system should match the maximum to it. The resulting file contains the relevance ranking of the documents generated by the IR system being evaluated. This file is evaluated by trec_eval based on the "correct answer" provided by the first file. The final field score is an integer or floating-point value that indicates the similarity between a document and a query, so the most relevant document has the highest score. Different measurement techniques can be used to verify the results, and in this study, we opted for precision at the top 5, 10, 15, 20, 30, and 100.

### *2.2. Pre-Trained Text Classification*

Text document classification is a classic problem in natural language processing (NLP). Historically, various neural models were used to learn text representations, including convolutional models such as sentences model [7], character level [8], very deep convolutional neural network (CNN) [9], and deep pyramid CNN [10]. Recurrent models [11] and attention mechanism models [12] were also studied. However, recent research conducted reveals that pre-trained models for large populations show excellent performance in text classification and multiple NLP tasks, without having to train new models from scratch [13]. Using large amounts of unlabeled data, OpenAI GPT (Open Artificial Intelligence Generative Pre-Training) [14] and BERT [15], show that pre-trained language models can help learn common language expressions.

Language models have a history, and several studies have discussed various language models. For instance, the n-gram language model [16] assumes that the current word can be predicted via the previous n words. The introduction of feedforward neural networks-based language models was advantageous; however, they suffer from pre-defined contexts [17]. The recurrent neural network to express a language and can theoretically use any length of context [18]. Recently, deep expression learning models such as BERT and Embedding from Language Models (ELMo) have been demonstrated to improve many NLP tasks [15]. These models typically learn language expressions from large raw text using unsupervised pre-training techniques [19]. Recent research has found that many downstream applications can benefit from the representation of words generated by pre-trained models [15]. The ELMo uses a bidirectional iterative neural network to create word expressions [20]. The BERT [15] is based on a multi-layer bidirectional Transformer [21] and uses two pre-training goals: the language model of the mask and the prediction of the next sentence, which allows getting natural benefits from large, unlabeled data. The BERT input consists of three parts: word piece, location, and segment. It uses a bidirectional converter to generate the word expressions which is co-adjusted in the left and right context of all layers. Derivatives such as BioBERT [22] recognize various NLP or biomedical NLP operations (e.g., question answering, specified objects, relationship extraction, document extraction through simple fine-tuning techniques).

### *2.3. Contextualized Word Representations*

The traditional pre-trained word representation is somewhat less comprehensible in language because of the notion of embedding the same word always in the same vector [23], while the same word can be used differently depending on the context. For instance, the popular word2vec assumes that the meanings of words that appear in similar locations are similar. It stores the meaning of the context when the word is vectorized. The words changed into vectors can be measured with "existing radian distances" such as "cosine similarity", and are interpreted as words with similar meaning when the distances are close together. The embedding is done on a word-by-word basis. In other words, the word vector is obtained by multiplying the one-hot-encoding vector of each word by WT [24]. On the counterpart, we can think that the same word can be used differently depending on the context. Therefore, the contextualized word embedding follows the assumption that the performance of natural language processing may increase if the same word is embedded differently according to the context. The contextualized word representation uses a very deep neural network to grasp the meaning of words according to the context, thus before embedding a word, it embeds the whole sentence.

### *2.4. Contextualized Word Embedding for Information Retrieval*

Recently, there has been a great deal of work in designing ranking architectures that effectively score pairs of query documents and drive results [25]. A neural ranking model, Conv-KNRM [26], uses a convolutional neural network rather than a word representation, allowing the model to learn situational awareness representations from the local proximity. The language representation of the pre-trained context depends on the situation. Unlike more common pre-trained word vectors such as GloVe [27], which generates a word representation for each word in the vocabulary, this model generates a representation for each word based on the context of the sentence. A framework called DeepCT is developed using neural contextualized text representations of BERT that learns to map the contextualized word representations onto the target term weights in a supervised manner [28]. The DeepCT framework proposes DeepCT-Index that estimates the importance of each document term and DeepCT-Query, which estimates the importance of each query term. The retrieval models such as BM25 and QL can use the new index and the new query directly for the retrieval tasks. A multitask attention-based bidirectional LSTM–CRF (Att-biLSTM–CRF) model is developed using pre-trained Embeddings from Language Models (ELMo) with enhanced functionality of named entity discovery to enrich the model's perception of unknown entities [29].

### 2.5. Self-Attention Model

Self-attention mechanisms have gained increased popularity in neural network attention research and demonstrated fruitful results in a wide variety of tasks, particularly natural language processing (NLP) tasks. The Google machine translation team first introduced it in a work titled "attention is all you need" by presenting the idea of Transformer [21], which was entirely based on attention, using multi-headed self-attention as a replacement of the recurrent layers most commonly used in encoder-decoder architectures. BERT, which is currently attracting attention as the best NLP model, also uses the bidirectional training of Transformer. Unlike the recurrent neural network (RNN), the Transformer, a model to solve the problem of entire sequence dependency, can perform the parallel calculation. For the input sequence, it finishes the attention calculation in parallel before calculating the output. Then, when calculating the output word, the next word is predicted using the attention of the words before the word to be output and the attention of the precomputed input.

### 2.6. Scoring Mechanisms for Information Retrieval

In a general document search engine, calculating a score value is called relevance and can be translated as accuracy. The TF-IDF (Term Frequency Inverse Document Frequency) and BM25 (Best Matching 25) are the most used score algorithms in search engines. TF-IDF is a weight used in information retrieval and text mining, and it works by comparing the relative frequency of a word in a document with the inverse ratio of that word across the document corpus. Intuitively, this calculation determines how related a particular word is to a specific document. Common words in single or few documents tend to have higher TF-IDF numbers than common words such as articles and prepositions [30]. Okapi BM25 is a scoring algorithm used in search engines and recommendation systems [31]. Like TF-IDF, the BM25 algorithm is also based on the concept of term frequency and inversed document frequency. However, it considers the length of the document in scoring.

Different techniques, as mentioned in the above paragraphs, are historically used by different researchers for a variety of tasks related to the relevant document retrieval to answer the user queries [1–3,30,32–37]. Our job is to investigate the best among the state-of-the-art techniques to use for the task of finding the high impact relevant documents and rank them to satisfy the user needs asked in the query.

## 3. Materials and Methods

The proposed approach of identifying highly relevant clinical articles matched with the given topic of interest is divided into two steps. In the first step, the data is processed to extract the title and abstract of each article, which are then used as features in the disease classification task. The output of this step is the subsets of articles where each subset refers to a disease. In the second step, a query is made from the topic by including only gene and demographic information, which is then matched with the title and abstract collected from a disease-specific subset of articles. At this stage, a score related to the gene importance in a document is also determined. Finally, the scores are combined to determine the final ranking of articles with the topic of interest. The sub-components of the proposed method are described in Figure 1.

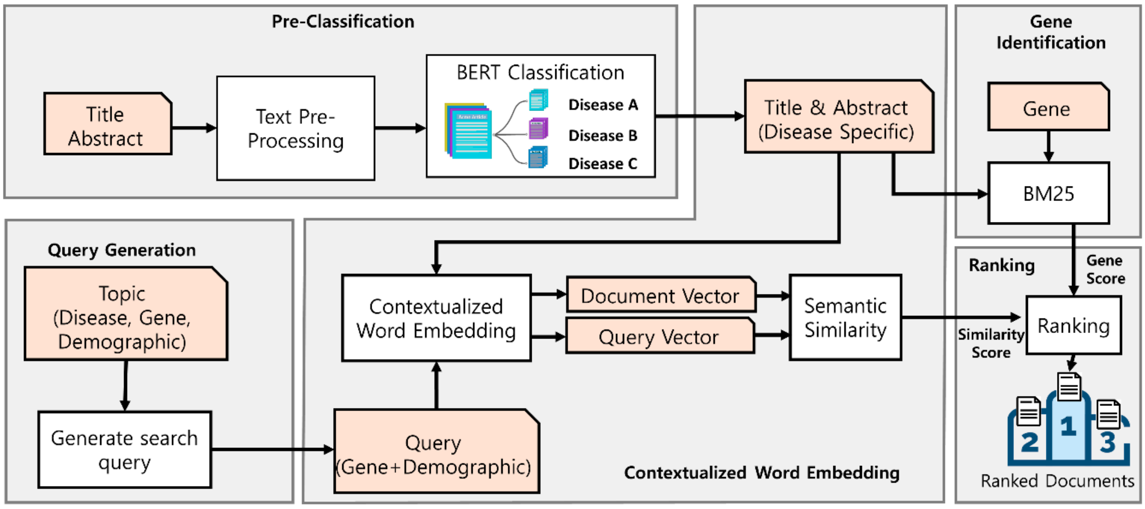

**Figure 1.** Proposed methodology of identification of relevant documents to satisfy a query generated from given topic using attention-based Bidirectional Encoder Representation from Transformers (BERT) classification, contextualized word embedding, and semantic similarity.

In the proposed method of finding relevant documents needed for precision medicine, first, the documents were classified into disease groups using a deep learning classifier to consider the disease. Second, a query is created from a given topic, which contains information about genes and demographics. The disease information is not included in the query as pre-classification has already classified the documents into different health conditions. Third, vectors are generated for the document (title and abstract) as well as the query using contextualized word embedding to calculate the similarity score. Fourth, the search is carried out for the gene information in the title and the abstract of documents using the BM25 algorithm, which helps to indicate how important the information about the gene is in the entire document. Finally, the documents are ranked based on the combined relevance score obtained from the relevance between query and document and the gene importance in a document.

### 3.1. Data Acquisition

We utilized TREC's 2018 Precision Medicine track data to rank relevant documents that matched the query. The medical document data provided by TREC consists of 240,324 biomedical documents (abstracts and titles) along with other meta information such as disease, author, document number, publication year, and journal name in XML format. First, we extracted the titles, abstracts, and disease labels from the data. In the precision medicine track, a collection of topics is provided where each topic comprises of three key elements in a semi-structured form: disease (e.g., type of cancer), gene variants (mainly gene variant of the tumor as opposed to patient DNA), and demographic information (age, sex).

### 3.2. Pre-Classification for Disease

For the classification, titles and abstracts in medical documents related to five health conditions that include breast cancer, healthy, HIV, melanoma, and prostate cancer were extracted. Initially, four prevalent deep learning methods were applied to find a suitable candidate for the classification task: CNN, RNN, Bidirectional Long Short-Term Memory (Bi-LSTM), and BERT. These algorithms are chosen based on their popularity for NLP tasks in recent times. The CNN algorithm has the advantage of layers using convolution filters applied to local features, which is originally devised for computer vision; however, it has shown to be effective for NLP and has achieved excellent results over traditional NLP algorithms [38]. The RNN algorithm is preferred for NLP tasks due to its recursive structure, which is suitable for variable length text processing. It can also take advantage of a distributed representation of words by first converting the tokens that make up each text into a vector that forms a matrix which is more suitable to capture as much contextual information as possible. It also uses a maximum pooling layer to automatically determine which words play an important role in text classification to capture important components of the text [39,40]. The attention-based BiLSTM algorithm captures the most important semantic information in a sentence using a bidirectional long short-term memory network. This model does not use lexical resources or features derived from NLP systems. It focuses on words that have a decisive impact on classification and captures the most important semantic information of a sentence without using additional knowledge or the NLP system [41]. The BERT base model has 12 transformer blocks, 12 self-attention heads, and an encoder with a hidden size of 768. Based on this model, BERT achieved state-of-the-art performance in a variety of downstream tasks. The input representation can represent both a single sentence and a set of sentences in one token sequence so that BERT can handle various downstream tasks. A "sentence" is an arbitrary range of continuous text, not an actual language sentence, which may be a single sentence, or two sentences packed together [13,15].

In this paper, we use sentence classification among General Language Understanding Evaluation (GLUE) tasks. Among several BERT fine-tuning for NLP tasks, classification tasks are sentence pair classification and single sentence classification tasks. The sentence pair classification, as described in [15], is employed to measure the performance of document classification. The sentence transformers' BERT framework is utilized to generate a semantically meaningful sentence embedding for document matching and ranking.

### *3.3. Query Creation*

A set of queries are created from the given set of topics. Of the three, two key elements are included in the query: demographics and gene information. The queries are created for a total of 25 topics provided by TREC. A partial list of queries is shown in Table 1.

**Table 1.** Queries for Precision Medicine Topics from Text REtrieval Conference (TREC) 2018.

| **Topic** | | | **Query** |
|---|---|---|---|
| **Disease** | **Gene** | **Demographic** | |
| Melanoma | BRAF (V600E) | 64-year-old male | BRAF V600E in old adult males |
| | BRAF (V600E), PTEN loss of function | 57-year-old male | BRAF V600E PTEN loss of function in old adult males |
| | KIT (L576P), KIT amplification | 56-year-old female | KIT L576P KIT amplification in old adult females |
| | no tumor-infiltrating lymphocytes | 74-year-old female | no tumor-infiltrating lymphocytes in old adult females |
| | high serum LDH levels | 69-year-old female | high serum LDH levels in old adult females |

The complete list of queries is provided in supplementary material Table A1.

### *3.4. Sentence Similarity with Contextualized Word Embedding*

Finding semantic similarity between textual passages is imperative for certain information retrieval tasks such as searching, query suggestions, automatic summarization, and image searching. Many approaches have been proposed based on lexical matching, handcrafted patterns, parse trees, external sources of structured semantic knowledge, and distribution semantics. However, lexical features such as string matching do not capture semantic similarity beyond a trivial level. In addition, external sources of handcrafted patterns and structured semantic knowledge cannot be assumed to be available in all situations and in all domains. Finally, parse tree-dependent approaches are limited to syntactically well-formed text, usually one sentence length [42]. Recent advances in neural language models have contributed to new ways of learning distributed vector representations of words. These methods have been shown to produce embedding that finds higher-order relationships between words that are highly effective for natural language processing tasks, including the use of word similarity and word inference [43]. Word embedding technologies such as word2vec are very enthralled in the NLP community that helps to get a list of words in the form of word vectors used in a similar context for a particular word [44]. Pre-trained word representation is an important component of many neural language understanding models. However, learning high-quality expressions can be difficult as it requires modeling of both (1) the complex characteristics of word usage and (2) how these usages vary with the context of the language. A new type of deep contextualized word representation has been introduced that solves both challenges of complex characteristics of word usages and variation of usages with the context of languages. Deep contextualized word representation vectorizes words contextually. In other words, the same word produces a different vector depending on the context. In this paper, we use the contextualized word representation for the title and abstract of medical documents to create word vectors. Additionally, the queries are vectorized with the same method. After getting the contextualized word vectors for titles, abstracts, and queries, we first determine the distance of a query with a title, then repeated the process for the abstract to get the individual similarity score of a query with the title as well as with the abstract. Finally, the individual scores are added to obtain the similarity score of a document with a given query.

### *3.5. Gene Importance Score*

The gene-related information is extracted from each topic and searched in documents to identify the importance of documents concerning that gene. Finding such information has an impact on the similarity score. Two methods are tested for gene information searching: TF-IDF and BM25. Because of the fundamental difference between the two methods concerning the length of a document, we found a significant difference in the score. Unlike TF-IDF, BM25 considers the average document length for calculation and ranks documents based on the query terms appearing in each document, regardless of their proximity within the document.

The individual scores found in Section 3.4 (similarity score of a query and a document) and Section 3.5 (gene importance score) are added. The resultant value is used as the ranking score based on which the documents are ranked and presented for the evaluation. The whole methodology of identifying similarity scores to rank the documents is given in Algorithm 1.

**Algorithm 1.** Algorithms to find document ranking based on similarity score aggregation.

```
Input:
D: The list of documents
G: Topic Gene Data
Q: Query
Output:
RD: Ranked list of documents according to their relevance score

Begin:
foreach title and abstract in D do:
    title_tokenization← tokenizer.tokenize(title)
    abstracts_tokenization ← tokenizer.tokenize(abstract)
endfor
title_embedding ← embedder.bertencoder(title_tokenization)
abstract_embedding ← embedder.bertencoder(abstracts_tokenization)
query_embedding ← embedder.bertencoder(Q)
title_bm25ranking ← bm25raking (title_tokenization)
abstract_bm25ranking ← bm25raking (abstracts_tokenization)
title_bm25_score ← title_bm25ranking.get_scores(tokenizer.tokenize(G)) *2
abstract_bm25_score ← abstract_bm25ranking.get_scores(tokenizer.tokenize(G)) *2
total_bm25_Score ← title_bm25_score + abstract_bm25_score
embedding_score_title ← sum (query_embedding * title_embedding)/title_embedding
embedding_score_abstract ← sum (query_embedding *abstract_embedding)/abstract_embedding
top_doc_ids ← get_top(embedding_score_title + embedding_score_abstract + total_bm25_Score)
foreach id in top_doc_ids do:
    score ← (embedding_score_title[id] + embedding_score_abstract[id] +total_bm25_Score[id])
    RD ← get (D[”id”], D[”Title”], score)
endfor
return RD
End
```

## 4. Results

To test and evaluate the proposed methods, a document collection of 2018 TREC Precision Medicine Track was utilized as well as the TREC system for the evaluation to rank documents.

### 4.1. Experiment Design

The experiment is performed in two stages as shown in Figure 2: (1) pre-classification stage and (2) document ranking. In the first stage, we tested the proposed BERT classifier to obtain a disease-specific set of documents making us able to conduct a search through a query only within a subset of documents. In the second stage, we obtained the results of ranking documents through similarity scores using query similarity with documents and gene importance.

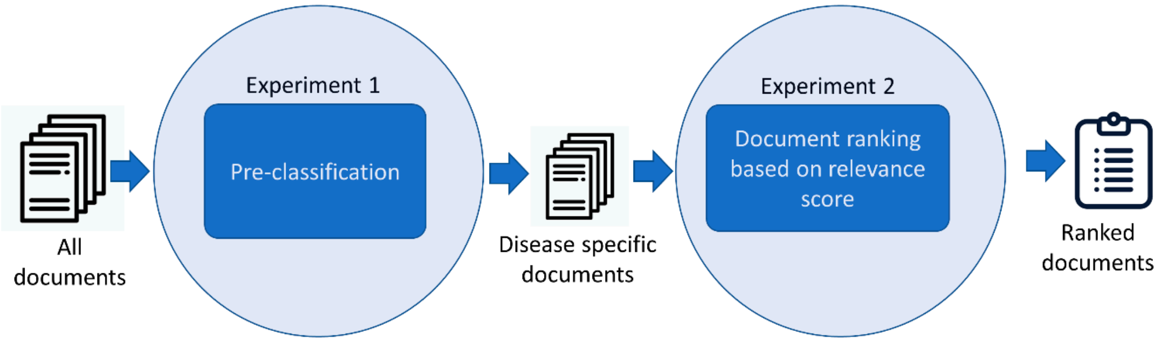

**Figure 2.** Experiment design where experiment 1 is designed for pre-classification and experiment 2 shows the document ranking based on relevance.

### *4.2. Pre-Classification Results and evaluation*

The five health conditions that include breast cancer, healthy, HIV, melanoma, and prostate cancer, and the number of documents distributed into the training and testing sets are shown in Table 2.

**Table 2.** Dataset of health conditions with training and testing distribution.

| Health Condition | Train | Test |
|---|---|---|
| Breast Cancer | 4717 | 1179 |
| Healthy | 4918 | 1230 |
| HIV | 3945 | 986 |
| Melanoma | 1222 | 306 |
| Prostate Cancer | 2266 | 566 |

A total of 21,335 medical document data where each document was composed of a title and an abstract was used for model development and testing. The data is divided at a ratio of 80% (train data) and 20% (test data). The performance of classifying documents of three deep learning models; CNN, RNN, and Bi-LSTM are compared with the proposed BERT classifier and their results are reported in Table 3. The BERT classifier has the highest score with improvements of about 1% (in accuracy and precision) and of about 8% (in training time) over the next best performer Bi-LSTM. Though the training time of CNN is substantially better than all classifiers on the list, it has the lowest metrics measured here. As BERT attained the highest precision out of the models tested, we chose to utilize it for the task of retrieving relevant documents.

**Table 3.** Performance comparison of four classifiers: convolutional neural network (CNN), recurrent neural network (RNN), Bidirectional Long Short-Term Memory (Bi-LSTM), and BERT.

| Classifier | Precision | Recall | f1-Score | Accuracy | Training Time (min) |
|---|---|---|---|---|---|
| BERT | 0.96 | 0.95 | 0.95 | 0.95 | 1158.22 |
| Bi-LSTM | 0.95 | 0.95 | 0.95 | 0.94 | 1252.93 |
| RNN | 0.94 | 0.94 | 0.94 | 0.94 | 1284.69 |
| CNN | 0.93 | 0.93 | 0.93 | 0.93 | 62.32 |

### *4.3. Relevance Document Ranking*

After getting the disease-specific documents, we proceeded to rank the documents by calculating the relevance scores using semantic similarity method. The ranked documents are then evaluated using the document ranking evaluation program provided by TREC. For cross-comparison, we demonstrated the results of two alternatives:

1. Relevance ranking results without pre-classification of health condition
2. Relevance ranking results with pre-classification of health condition

In each of the above scenarios, we compare the results of searching in four runs. In the second scenario, i.e., with pre-classification, we get comparatively better results. Figure 3 shows the box-and-whisker plot for the two scenarios. The top-performing run precision score was 0.512 at P@5 for the second scenario, which is about 1% better than the score of the first scenario.

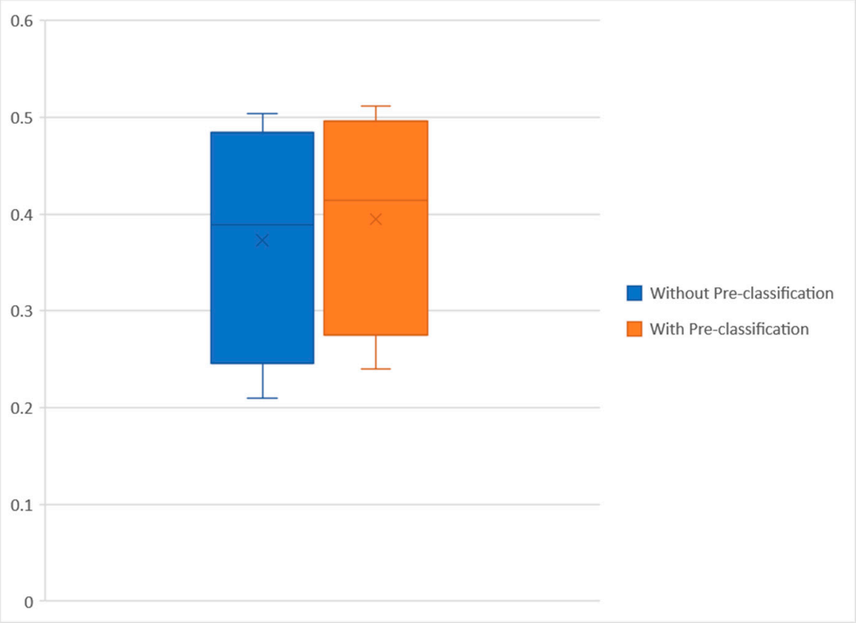

**Figure 3.** Scores of four runs for two scenarios; with and without pre-classification.

Details of the four runs in each scenario are described in Tables 4 and 5. Table 4 shows the results of searching for all medical documents for the first scenario.

**Table 4.** Relevance ranking results at different precision levels (without pre-classification).

| Method | P@5 | P@10 | P@15 | P@20 | P@30 | P@100 |
|---|---|---|---|---|---|---|
| CWE (Query) | 0.2100 | 0.1620 | 0.1253 | 0.1140 | 0.1060 | 0.0676 |
| CWE (Query)<br>TF-IDF (Gene) | 0.3520 | 0.3500 | 0.2986 | 0.2548 | 0.1645 | 0.1150 |
| BM25 (Query)<br>BM25 (Gene) | 0.4260 | 0.3600 | 0.3040 | 0.2860 | 0.2400 | 0.1380 |
| CWE (Query) and BM25 (Gene) | 0.5040 | 0.4200 | 0.3680 | 0.3260 | 0.2773 | 0.1824 |

**Table 5.** Relevance ranking results at different precision levels (with pre-classification).

| Method | P@5 | P@10 | P@15 | P@20 | P@30 | P@100 |
|---|---|---|---|---|---|---|
| CWE (Query) | 0.2400 | 0.1800 | 0.1400 | 0.1300 | 0.1100 | 0.0850 |
| CWE (Query)<br>TF-IDF (Gene) | 0.3800 | 0.3900 | 0.3267 | 0.2700 | 0.2067 | 0.1390 |
| BM25 (Query)<br>BM25 (Gene) | 0.4480 | 0.3880 | 0.3387 | 0.3020 | 0.2653 | 0.1800 |
| Proposed (CWE (Query) and BM25 (Gene) | 0.5120 | 0.4480 | 0.3840 | 0.3440 | 0.2893 | 0.1964 |

Run 1: the search was performed using a sentence similarity between a generated query and a medical document using the contextualized word embedding (CWE) method. Run 2: the CWE (query) is combined with the TF-IDF-based gene to get the results. Run 3: the BM25 (query) search is combined with the BM25 (gene) for searching. Run 4: the CWE (query) is combined with the BM25 (gene) to find the relevant documents. In all these runs, we measure the precision of the top five, ten, fifteen, twenty, thirty, and one hundred of the medical documents related to the topic. The fourth run showed the highest precision of 0.50 at P@5, which is about 16% better than the next best performer.

For the second scenario, i.e., the relevance ranking results with pre-classification, we searched only within the disease-specific subset of documents. The experiment was conducted the same way as for the first scenario, except the pre-classification. The purpose was to check the benefit of the pre-classification on the results. As shown in Table 5, the pre-classification adds a slight improvement (from 1 to 4%) to the results at different precision levels. For instance, the best run gets about 4% better accurate results at P@10 on the pre-classified data than the data without pre-classification.

## 5. Conclusions

In this paper, we propose a method for retrieving the medical evidence documents most relevant to the topic of interest. The proposed method first classifies the data based on a health condition using the BERT classifier. Using contextualized word embedding, we obtain contextually-enriched vectors of the query and the medical documents to find the similarity score. We then identify the importance score of the genetic information in documents. Finally, we get the relevance ranking score by adding the similarity score and gene importance score. The proposed method is evaluated on data of 2018 TREC's Precision Medicine Track. In the pre-classification, we get the classification accuracy based on which we obtain the disease-specific set of documents. Within the disease-specific documents, we rank the documents by measuring the relevance to the topic. We can observe in the results obtained through the TREC evaluation program that the proposed method obtained better results than the baseline, which provides the confidence to use it for the retrieval tasks in other domains.

Our proposed method achieves comparatively better results for two reasons: pre-classification before performing query-based searching and the add-on of gene importance score to the similarity score of query and document. The pre-classification segregates the unnecessary documents and forwards a focused set for further searching. Without the pre-classification, the chances of getting unnecessary documents (false positives) may increase, i.e., the query may retrieve documents that have disease information, not as the main subject rather as a secondary focus. For instance, a document about cancer with a focus on breast cancer, and thyroid cancer is just used as an example somewhere in the body of the text. Without pre-classification, there is a higher chance for a query created to retrieve thyroid cancer documents along with breast cancer documents. The gene importance score brings those documents to the top that have the gene as their main subject. In other words, it adds up to the overall relevance ranking score. Therefore, our proposed method with pre-classification and the gene importance score provides a different way to find and rank relevant documents.

Dealing with parallelism and long-range dependencies was a big challenge where the transformer framework encounters this challenge by introducing a self-attention model that calculates attention multiple times in parallel and independently. The self-attention layer in the BERT model has overcome the challenge of long-range dependencies and therefore gets comparatively better results than conventional RNN and Bi-LSTM models. Specifically, the BERT classifier performed better than the competitors in terms of accuracy (95%) and precision (96%); therefore, we opted to use it as our proposed model in this work to obtain the more precise results. The Python code of the proposed is made available on a publicly available repository on GitHub (https://github.com/jamilbadama/Ranking_Trec_Documents_Deeplearning), which can be reused for other datasets. For instance, our models of pre-classification and ranking can be reused for the datasets 2019 TREC Precision Medicine topic and onward.

Despite the fact that we have achieved comparatively precise results; however, this is on the cost of missing about 5% documents during the pre-classification stage as our proposed model accuracy was 95%. Improving the pre-classification accuracy will reduce the chances of missing documents to use for query matching in the later steps. Furthermore, this experiment is performed on the 2018 TREC dataset, and since the 2019 dataset is now available, it will be useful to test the proposed model on the new dataset with more refinements to the models.

In conclusion, the contextually viable and competitive outcomes of the proposed model confirm the suitability of our proposed model for use in various domains where clinical studies are part of a clinical care, such as precision medicine, evidence-based medicine, and medical education.

**Author Contributions:** B.P. and M.A. are the principal researchers and the main authors of this work. They conceptualized the main idea of the work. B.P did research and implement the proposed research methods and algorithms. M.A refined the methodology and added further contents to the main body of the text. J.H. devised thematic aspects of the workflow and improved visualization and also improved the implementation code. A.A. helped in data preparation. S.L. supervised the work and acquired funding for the project. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2017-0-01629) supervised by the IITP (Institute for Information and communications Technology Promotion). This work was supported by the Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2017-0-00655), NRF-2016K1A3A7A03951968, and NRF-2019R1A2C2090504.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A

**Table A1.** Queries for Precision Medicine Topics from TREC 2018.

| Topic | | | Query |
|---|---|---|---|
| **Disease** | **Gene** | **Demographic** | |
| Melanoma | BRAF (V600E) | 64-year-old male | BRAF V600E in old adult males |
| Melanoma | BRAF (V600K) | 54-year-old male | BRAF V600K in middle adult males |
| Melanoma | BRAF (V600R) | 80-year-old male | BRAF V600R in middle old males |
| Melanoma | BRAF (K601E) | 38-year-old male | BRAF K601E in adult males |
| Melanoma | BRAF (V600E), PTEN loss of function | 57-year-old male | BRAF V600E PTEN loss of function in old adult males |
| Melanoma | BRAF (V600E), NRAS (Q61R) | 67-year-old male | BRAF V600E NRAS Q61R in old males |
| Melanoma | BRAF amplification | 61-year-old male | BRAF amplification in old adult males |
| Melanoma | NRAS (Q61R) | 63-year-old female | NRAS Q61R in old adult females |
| Melanoma | NRAS (Q61L) | 34-year-old female | NRAS Q61L in adult females |
| Melanoma | KIT (L576P) | 65-year-old female | KIT L576P in old adult females |
| Melanoma | KIT (L576P), KIT amplification | 56-year-old female | KIT L576P KIT amplification in old adult females |
| Melanoma | KIT (K642E) | 62-year-old female | KIT K642E in old adult females |
| Melanoma | KIT (N822Y) | 39-year-old female | KIT N822Y in adult females |
| Melanoma | KIT amplification | 66-year-old female | KIT amplification in old adult females |
| Melanoma | NF1 truncation | 70-year-old male | NF1 truncation in old adult males |
| Melanoma | NTRK1 rearrangement | 60-year-old male | NTRK1 rearrangement in old adult males |
| Melanoma | TP53 loss of function | 72-year-old male | TP53 loss of function in old adult males |
| Melanoma | tumor cells with >50% membranous PD-L1 expression | 48-year-old female | tumor cells with >50% membranous PD-L1 expression in adult females |
| Melanoma | tumor cells negative for PD-L1 expression | 73-year-old male | tumor cells negative for PD-L1 expression in old adult males |
| Melanoma | high tumor mutational burden | 86-year-old female | high tumor mutational burden in old adult males |

**Table A1.** *Cont.*

| Topic | | | Query |
|---|---|---|---|
| **Disease** | **Gene** | **Demographic** | |
| Melanoma | extensive tumor infiltrating lymphocytes | 49-year-old male | extensive tumor infiltrating lymphocytes in adult males |
| Melanoma | no tumor infiltrating lymphocytes | 74-year-old female | no tumor infiltrating lymphocytes in old adult females |
| Melanoma | PTEN loss of function | 68-year-old male | PTEN loss of function in old adult males |
| Melanoma | APC loss of function | 47-year-old male | APC loss of function in adult males |
| Melanoma | high serum LDH levels | 69-year-old female | high serum LDH levels in old adult females |

## References

1. Bian, J.; Abdelrahman, S.; Shi, J.; Del Fiol, G. Automatic identification of recent high impact clinical articles in PubMed to support clinical decision making using time-agnostic features. *J. Biomed. Inform.* **2019**, *89*, 1–10. [CrossRef] [PubMed]
2. Bian, J.; Morid, M.A.; Jonnalagadda, S.; Luo, G.; Del Fiol, G. Automatic identification of high impact articles in PubMed to support clinical decision making. *J. Biomed. Inform.* **2017**, *73*, 95–103. [CrossRef] [PubMed]
3. Afzal, M.; Hussain, M.; Malik, K.M.; Lee, S. Undefined Impact of Automatic Query Generation and Quality Recognition Using Deep Learning to Curate Evidence from Biomedical Literature: Empirical Study. *JMIR Med. Inform.* **2019**, *7*. [CrossRef] [PubMed]
4. MacAvaney, S.; Cohan, A.; Yates, A.; Goharian, N. CEDR: Contextualized embeddings for document ranking. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 1101–1104.
5. Text REtrieval Conference (TREC) Overview. Available online: https://trec.nist.gov/overview.html (accessed on 19 February 2020).
6. GitHub—Usnistgov/trec_eval: Evaluation Software Used in the Text Retrieval Conference. Available online: https://github.com/usnistgov/trec_eval (accessed on 20 February 2020).
7. Kalchbrenner, N.; Grefenstette, E.; Blunsom, P. A convolutional neural network for modelling sentences. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22 June 2014; Volume 1, pp. 655–665.
8. Zhang, X.; LeCun, Y. Text Understanding from Scratch. *Adv. Neural Inf. Process. Syst.* **2015**, 649–657.
9. Duque, A.B.; Santos, L.L.J.; Macêdo, D.; Zanchettin, C. Squeezed very deep convolutional neural networks for text classification. In Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Munich, Germany, 17–19 September 2019; Volume 11727, pp. 193–207.
10. Johnson, R.; Zhang, T. Deep pyramid convolutional neural networks for text categorization. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; Volume 1, pp. 562–570.
11. Yogatama, D.; Dyer, C.; Ling, W.; Blunsom, P. Generative and Discriminative Text Classification with Recurrent Neural Networks. *arXiv* **2017**, arXiv:1703.01898.
12. Lin, Z.; Feng, M.; Dos Santos, C.N.; Yu, M.; Xiang, B.; Zhou, B.; Bengio, Y. A structured self-attentive sentence embedding. In Proceedings of the 5th International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
13. Sun, C.; Qiu, X.; Xu, Y.; Huang, X. How to fine-tune BERT for text classification? In Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Berlin, Germany, 13 June 2019; Volume 11856, pp. 194–206.
14. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 21 August 2020).

15. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 8 June 2019; Volume 1, pp. 4171–4186.
16. Manning, C.D.; Schütze, H.; Weikurn, G. Foundations of Statistical Natural Language Processing. *SIGMOD Rec.* **2002**, *31*, 37–38.
17. Sundermeyer, M.; Ney, H.; Schluter, R. From feedforward to recurrent LSTM neural networks for language modeling. *IEEE Trans. Audio Speech Lang. Process.* **2015**, *23*, 517–529. [CrossRef]
18. Mikolov, T.; Kombrink, S.; Burget, L.; Černocký, J.; Khudanpur, S. Extensions of recurrent neural network language model. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Prague, Czech Republic, 22–27 May 2011; pp. 5528–5531.
19. Li, F.; Jin, Y.; Liu, W.; Rawat, B.P.S.; Cai, P.; Yu, H. Undefined Fine-Tuning Bidirectional Encoder Representations from Transformers (BERT)–Based Models on Large-Scale Electronic Health Record Notes: An Empirical. *JMIR Med. Inform.* **2019**, *7*, e14830. [CrossRef] [PubMed]
20. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; Volume 1, pp. 2227–2237.
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 2017, pp. 5999–6009.
22. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2019**. [CrossRef] [PubMed]
23. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. In Proceedings of the 1st International Conference on Learning Representations, Scottsdale, AZ, USA, 2–4 May 2013.
24. Contextualized Word Embedding (Concept)—Woosung Choi's Blog. Available online: http://intelligence.korea.ac.kr/members/wschoi/nlp/deeplearning/paperreview/Contextualized-Word-Embedding/ (accessed on 24 February 2020).
25. Guo, J.; Fan, Y.; Ai, Q.; Croft, W.B. A deep relevance matching model for Ad-hoc retrieval. In Proceedings of the International Conference on Information and Knowledge Management, Indianapolis, IN, USA, 26–28 October 2016; pp. 55–64.
26. Dai, Z.; Callan, J.; Xiong, C.; Liu, Z. Convolutional neural networks for soft-matching N-grams in ad-hoc search. In Proceedings of the 11th ACM International Conference on Web Search and Data Mining, Marina Del Rey, CA, USA, 5–9 February 2018; Volume 2018, pp. 126–134.
27. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
28. Dai, Z.; Callan, J. Deeper text understanding for IR with contextual neural language modeling. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 985–988.
29. Yang, J.; Liu, Y.; Qian, M.; Guan, C.; Yuan, X. Information extraction from electronic medical records using multitask recurrent neural network with contextual word embedding. *Appl. Sci.* **2019**, *9*, 3658. [CrossRef]
30. Juan Ramos Using tf-idf to determine word relevance in document queries. *Proc. First Instr. Conf. Mach. Learn.* **2003**, *242*, 29–48.
31. Robertson, S.; Zaragoza, H. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.* **2009**, *3*, 333–389. [CrossRef]
32. Oleynik, M.; Faessler, E.; Sasso, A.M.; Kappattanavar, A.; Bergner, B.; Freitas Da Cruz, H.; Sachs, J.-P.; Datta, S.; Böttinger, E. *HPI-DHC at TREC 2018 Precision Medicine Track*; TREC: Austin, TX, USA, 2018.
33. Pasche, E.; Van Rijen, P.; Gobeill, J.; Mottaz, A.; Mottin, L.; Ruch, P. SIB text mining at TREC 2018 precision medicine track. In Proceedings of the TREC 2018 Conference, Gaithersburg, MD, USA, 14–16 November 2018.

34. Ronzano, F.; Centeno, E.; Pérez-Granado, J.; Furlong, L. *IBI at TREC 2018: Precision Medicine Track Notebook Paper*; TREC: Austin, TX, USA, 2018.
35. Taylor, S.J.; Goodwin, T.R.; Harabagiu, S.M. *UTD HLTRI at TREC 2018: Precision Medicine Track*; TREC: Austin, TX, USA, 2018.
36. Zheng, Z.; Li, C.; He, B.; Xu, J. *UCAS at TREC-2018 Precision Medicine Track*; TREC: Austin, TX, USA, 2018.
37. Zhou, X.; Chen, X.; Song, J.; Zhao, G.; Wu, J. *Team Cat-Garfield at TREC 2018 Precision Medicine Track*; TREC: Austin, TX, USA, 2018.
38. Lopez, M.M.; Kalita, J. Deep Learning applied to NLP. *arXiv* **2017**, arXiv:1703.03091.
39. Zhou, P.; Qi, Z.; Zheng, S.; Xu, J.; Bao, H.; Xu, B. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. In Proceedings of the 26th International Conference on Computational Linguistics, Osaka, Japan, 11–16 December 2016; pp. 3485–3495.
40. Lai, S.; Xu, L.; Liu, K.; Zhao, J. Recurrent convolutional neural networks for text classification. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.
41. Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; Xu, B. Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 207–212.
42. Kenter, T.; De Rijke, M. Short text similarity with word embeddings. In Proceedings of the International Conference on Information and Knowledge Management, Melbourne, Australia, 19–23 October 2015; pp. 1411–1420.
43. Zuccon, G.; Koopman, B.; Bruza, P.; Azzopardi, L. Integrating and evaluating neural word embeddings in information retrieval. In Proceedings of the ACM International Conference Proceeding Series, Parramatta, Australia, 8–9 December 2015; pp. 1–8.
44. Ganguly, D.; Roy, D.; Mitra, M.; Jones, G.J.F. A word embedding based generalized language model for information retrieval. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, 9–13 August 2015; pp. 795–798.