Graph-based Similarity for Document Retrieval in the Biomedical Domain

Adelaida, A, Zuluaga
Department of Electrical and Electronic Engineering.
Universidad de los Andes, Colombia
a.zuluaga10@uniandes.edu.co

Andres, A, Rosso
School of Exact Sciences and Engineering. Universidad
Sergio Arboleda, Colombia
andres.rosso@usa.edu.co

ABSTRACT

The growing amount of available data in the biomedical domain turns out to be beneficial for decision-making, but a sufficiently accurate DR system is required. Plenty of NLP techniques and models have been proposed for semantic similarity in DR, but few of them have been able to consider the variations of the language and relationship between distant words in texts. This work is focused on formulating a Graph-based Similarity for DR method (GBS-DR) for the biomedical domain and comparing the obtained results with traditional DR paradigms. The graph-based methods were selected to prove the importance of analyzing the semantic, syntactic, and long-distant word relationships in texts. It will be demonstrated that through the graph's topology the system can extract the structural information of documents, which solves relevant issues that are faced in this research area.

CCS CONCEPTS •Information Systems •Information Retrieval •Retrieval Models and Ranking •Learning to Rank

KEYWORDS

Natural Language Processing, Document Retrieval, Search Engines, Biomedical Literature, Graphs

ACM Reference Format:
Adelaida, A, Zuluaga and Andres, A, Rosso. 2022. Graph-based Similarity for Document Retrieval in the Biomedical Domain. In *2022 7th International Conference on Machine Learning Technologies (ICMLT) (ICMLT 2022), March 11–13, 2022, Rome, Italy.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3529399.3529428

1 INTRODUCTION

The increasing number of published articles in the biomedical field, inevitably leads to the increasing need for a better information access system. In the biomedical domain, the goal of information access systems is to improve the gathering of valuable information for users who have questions about health-related issues. More precisely, there is a need to strictly accessing to useful and pertinent information regarding the user's query. DR is an information access paradigm which allow users to access relevant documents reducing the manual validation effort. Based on a biomedical posed query using natural language, a DR system must look for the most relevant documents that would offer the proper and accurate information. This paradigm consists of many stages, from the formulation of the question, through the representation of the documents, to the ranking and retrieval of them.

DR methods have mainly explored traditional Natural Language Processing techniques, that seek to find semantic and grammatical structure similarity between queries and documents. Thus, some of the most popular techniques, like word-embeddings (e.g., Word2Vec), evaluate the relationship between words in restricted window sizes, which means that the context captured for words is mainly based on the 2n surrounding words, these are known as sequential word algorithms. Other strategies, as those based on graph representations, provides additional information, since they sought to offer a different approximation that considers a wider range of words used for capturing their context. Through the characteristics and attributes of a graph, the system can analyze the syntactic and semantic relationship between words, regardless of the distance between them. Therefore, the retrieval of documents is even more precise when this kind of model is integrated into the DR system.

In this regard, the general objective of this work is to develop and integrate a Graph-based Similarity DR method (GBS-DR). The proposed algorithm consists of an efficient graph-based representation method to encode queries and documents, followed by a deep neural model that uses the graph-based representation to perform an effective DR. Using biomedical questions and documents from the NFCorpus dataset [1], the proposed model was systematically validated and compared against traditional and state-of-the-art models. It is empirically proved the importance of considering a study based on the relationship between distant (i.e., semantically, and syntactically different words) and close words. With the integration of graph-based representations to the DR process, the system can capture a wider context between words, which allows the retrieval of documents to be much more accurate. As a result, the proposed method outperforms several baselines and some state-of-the-art methodologies

2 RELATED WORK AND BACKGROUND

2.1 Biomedical DR

The Biomedical domain has important challenges to consider, two of the strongest ones are the hardy and professional vocabulary, and the relations between concepts, which tends to be complicated since it can vary from being an explicit, implicit, direct, indirect, known, or unknown relation [2].

There are several methods that have been proposed to address the DR problem in the biomedical domain. The implemented algorithms differ in the way language is represented, processed, and

ICMLT 2022, March 11–13, 2022, Rome, Italy

ACM ISBN 978-1-4503-9574-8/22/03…$15.00
https://doi.org/10.1145/3529399.3529428

understood. In terms of the information representation, traditional strategies have been used several times, for instance Brokos et al. [3] represented documents and queries as the centroids of their word embeddings. Comparing with this strategy, other research works used more sophisticated approaches for the representation of information, which are focused on graph-based strategies. Zhao et al. [2] used document-concept graph construction, and Zhang et al. [4] represented documents as keyword graphs.

Other methods make use of vector space representation to take advantage of the latent semantic space generated. For instance, [1] computed the centroid of each question and each document, and then in terms of cosine similarity, the top k-nearest centroids were retrieved. Moreover, [2] used a Document-Concept Graph augmented recurrent neural network model (RNN) to learn the representation of information and then, a learning-to-rank model was proposed, to identify relevant biomedical articles for a given query.

2.2 Vector Space-based Models

One of the strategies commonly used for the information retrieval task is based on vector space models (VSM). These methods consist in a transformation of the tokens of the texts into a numerical format, which is known as word embeddings. There are different paths for vector representations, including term weighting schemes, the skip-gram model, Continuous Bag of Words (CBOW) [5], among many other. Nowadays there are pretrained word embeddings that are derived from language modelling, such as the Transformers [6], and Word2Vec [5]. Word2Vec is a model based on Attention Mechanisms which can be endorsed by Common Bag of Words (CBOW) or Skip-Gram. In both cases, the semantic of words is apprehended so the models can capture dependencies between words and their relative meaning**.**

With the vectorization of terms, the semantic and syntactic relation between word embeddings is estimated, to rank documents according to their relevancy degree. Most of the IR engines make use of TF-IDF. The BM25 approach is based upon the fundaments of the TF-IDF measure, which means that the rank assigned to a document is nothing but the consideration of the frequency of appearance of terms in each document, hence it doesn't consider the proximity of relation of words within the texts.

The problem with these kinds of approaches is the limitation of context, since the representation of phrases is achieved through the capture of words that are closed to each other in the phrase, which means that the sliding window for capturing context along the phrases, is limited. Hence, there are other approaches for vector space modelling that can map phrases considering wider contextual information. For instance, graph-based representations manage to deal with the problem of limited context, thanks to their structure and some characteristics that will be mentioned later.

2.3 Graph-based Representations

According to [7], document matching methods based on deep learning models can be categorized in two groups, the semantic matching, and the relevance matching. The relevance matching emphasizes on capturing relevance and interaction signals, or matching patterns, which is the case of the studies proposed by [2, 4, 8]. These studies share their central idea, which is to use graph-based methods. Since terms in queries are not always presented together, and sometimes not even close in the candidate documents, it is necessary to analyze long-distance word relationships. Thus, it is tactical to use a deep learning model based on graph representations, since it allows a more detailed analysis despite the distance between words. Modelling documents and queries through graph structures, grants the assimilation of the relationships between distant words, through their connections in the graph [7]. Recent research towards graphs for NLP, Yao et al. (2019) [9] and Zhang et al. (2020) [10] have proposed the use of graph neural networks as a language model and proved that these models are adequate for capturing long-distance word dependencies. There are lots of advantages when working with graph-based strategies, and the fact that they offer a flexible representation learning process for the queries and the documents is one of them. "The learned representations can encode the underlying semantics for queries and documents" [2].

Graph-based representations for documents and queries grants, not only the assimilation of the semantic between similar words, but also it can consider relationships between synonyms, between words with variation on their nomenclature, among others. To manage one of the most frequent issues presented with the vocabulary, which is the vocabulary mismatch, Jibril Frej et al. proposed the model Knowledge Based Transformer Model for Information Retrieval (KTReL) [8]. The vocabulary mismatch occurs due to the different extensions of texts, or because of the ambiguity or specialization of the domains in which queries and documents are built. With KTReL the following schemas were integrated: Knowledge Based models, a Neural-Learning-to-Rank architecture that uses embeddings, Knowledge bases and Transformer encoders for IR. The NLTR models integrated by Jibril Frej et al. use prior knowledge from word embeddings trained on large amounts of texts. Besides, the knowledge bases were used to expand queries and documents to introduce related concepts/entities to the relevant information in the texts. Lastly, transformers encoders were incorporated to associate sequences of word embeddings and concept embeddings to a fixed-size representation. [8]

3 GRAPH-BASED SIMILARITY MODEL FOR DOCUMENT RETRIEVAL (GBS-DR)

The proposed model GBS-DR, captures query-document semantic relevance in two stages, as can be seen in Figure 1. The first stage consists of the graph construction, while the second one is the whole learning and matching process carried out through a Graph Convolutional Neural Network (GCNN). This model takes advantage of a rich representation obtained using graph similarity, to overcome some of the representation challenges previously exposed.

3.1 Textual Representations

With the purpose of working only with the most relevant information from each document, an entity recognition process was executed for every single document that composes the corpus, using SciSpacy. Thereafter, the identified entities and the keywords

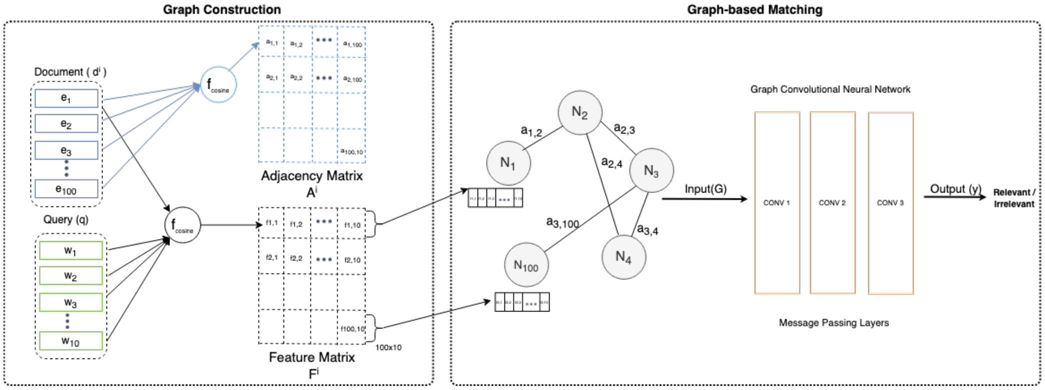

Figure 1: The architecture of the GBS-DR model that consists of two stages: graph construction and the graph-based matching, or the learning process @Adelaida Zuluaga

from the queries were transformed to their vectorial representation, obtaining the e_j and e_k as the embeddings of the entities and keywords, respectively. The embeddings were obtained using Word2Vec, and their dimension was set in 300.

The word embeddings are the basis for the construction of the graphs, because with these vectors the cosine similarity between entities and keywords is calculated, which makes up the feature and adjacency matrix. As it will be shown later, with these two matrices the topology of the graph is generated.

3.2 Graph Construction

The graph representation seeks to capture the interaction between documents and queries, which is known as interaction signal. To capture this relationship, for each query-document pair (q, di), a feature matrix F^i must be calculated, which is later transformed into the graph-of-words form; where d is an individual document and i ranges from 0 to the size of the corpus, and q corresponds to a particular query.

Considering that a graph is composed by nodes (N) and edges (E), the proposed topology consists of a set of nodes that is defined by the number of document entities, and the information of each one of them, or the node feature, is contained on a feature vector, found on F^i. The maximum number of nodes that was considered for each graph is 100, which stands for the maximum number of entities found in each document; this number is based on the distribution of entities for the complete corpus, so the set of nodes corresponds to the identified entities for each document, as shown in (1). In the case of the query terms, the maximum number of tokens was defined as 10, which is also based on the corpus distribution. In case the number of entities of the documents or the tokens contained in each query, are less than the values previously mentioned, zero padding was treated to fill in the missing information. This was done with the means of standardizing the sizes of the constructed graphs.

On the other hand, the edges will connect those nodes whose cosine similarity is greater than 0.6, which is defined by the similarity matrix A^i.

$$\{e_1, e_2, \ldots\ldots e_{100}\ \} \tag{1}$$

With the aim of finding the interaction signals between every entity-token pair, the feature matrix F^i was calculated. This matrix consists of the cosine similarities of the previously mentioned pair of words. This strategy was formulated just as Xueli Yu, et al. did in [7].

F^i represents the interaction signals, between the query and the retrieved d^i. The related matrix is denoted as $F^i\ R\ 100 \times 10$, where $f^i_{j,k}$ is the cosine similarity between the entity e_j, of the document d^i, and the keyword w_k of a query q. This value is formulated as presented in (2), where e_j and w_k q denote the word embeddings for e_j and w_k respectively. Thus, it is assigned to each node its corresponding feature vector, of size 1×10, contained on F^i.

$$s^i_{j,k}\ =\ cosine\left(e_j\ , w_k\ \right) \tag{2}$$

The last step in the graph representation process is connecting the correlated nodes. Each edge was created taking into consideration, for each pair of entities in d^i, the cosine similarity between their embeddings. Like so, the cosine similarity between $e_k d^i$ and $e_l d^i$ was calculated, building undirected edges between a pair of nodes if their cosine similarity was greater or equal to 0.6. On this account, the nodes connections, *node edges* constitute the adjacency matrix A^i, as shown in (3). With this done, the interaction between documents and queries have been transformed into an undirected graph. The next step is focused on finding relevance matching through a deep learning model.

$$A^i\ =\ \begin{cases} 1 & if\ cosine\left(e_k,\ e_l\right)\ >\ 0.6 \\ 0 & other\ wise \\ \end{cases} \tag{3}$$

3.3 Graph-based Matching

Having the graph G constructed, the process of graph-based matching is executed. This stage seeks to integrate a deep learning model that is structured based on the topology of graphs. Just as proposed in [4] and [7]. The proposed model consists in a deep relevance model based on multi-layer graph convolutional networks. As it can be seen in Figure 1, the previously constructed undirected graphs are the inputs for the Graph Convolutional Neural Network (GCNN). The convolutional layers will carry out a process called *neighborhood aggregation*, or *message passing* scheme. Finally, through the matching scores between nodes, a final tag y will be predicted, to determine if the input graph represents a relevant or irrelevant document for the given query.

Through a GCNN, the query-document interaction and the interaction between document entities are learned by *neighborhood aggregation*. Each node contains relevant information, in their corresponding feature vectors, and serves as bridge connecting with other nodes. “Through propagating the state representations to a node from its neighbors, it can receive the contextual information within the first-order connectivity"[7]. The number of times the aggregation occurs, is proportional to the number of information each node can receive from its neighbors. In other words, when aggregating t times, a node can receive information propagated from its *t-hop* neighbors. As it is explained in [7], through these aggregations the model achieves high-order aggregation of the query-document interactions, as well as the intra-document interaction. To incorporate contextual information into the word nodes, there must be an updating state. Through this, the representation of each node is updated.

4 EXPERIMENTAL SETUP

4.1 Dataset

The selected dataset is NFCorpus [1], which is publicly available for learning-to-rank in the medical domain. It consists of 5,276 different queries written in plain English and 3,633 documents composed of titles and abstracts from PubMed, with a highly technical vocabulary. The questions were written by biomedical experts, and they also identified the relevant documents, extracted from PubMed. The documents were normalized, tokenized, and stop words were removed. The queries and abstracts have an average length of approximately 13 and 143 tokens, respectively.

4.2 Baselines

4.2.1 BM25. Beforehand, there must be an initial selection of relevant documents, which is done through BM25. Using the indexed documents in Elasticsearch (ES), the optimal values for the parameters b and k1, of BM25, were searched. The obtained values are 0.8 and 0.1, respectively. With these parameters, the similarity between documents and queries was calculated and then, the search engine selected the top K documents, which are equivalent to the top relevant ones.

4.2.2 Centroid Embeddings for DR.. The Centroid Embedding (CEMB) for Document Retrieval consists of a Vector Space Representation model, which is basically the comparison between the centroids of the queries' embeddings and the document embeddings, just as it was proposed in [3]. CEMB takes 4 steps, the first step is the vectorization of documents and queries, which was made using Word2Vec, then the embedded documents word centroid is calculated as follows.

$$\vec{t} = \frac{\sum_{j=1}^{|v|} \overrightarrow{w_j} \cdot TF\left(w_j, t\right) \cdot IDF\left(w_j\right)}{\sum_{j=1}^{|v|} TF\left(w_j, t\right) \cdot IDF\left(w_j\right)} \tag{4}$$

Having each document represented by their centroid $\vec{t}$ as well as the centroid $\vec{q}$ of each query, the k top relevant documents are extracted based on the cosine similarity between them. The optimal number of documents k, based on MAP and RECALL metrics found in experimentation are 20.

4.3 Proposed Model Parameters

Following the strategy presented in Section 3, the graph representations were done in query-documents pairs. In first place, in terms of the pair samples per query, the maximum number of document-query pair, for each query, was defined to be 40. Meaning that, for each query at most 20 relevant documents, obtained from the top 20 documents retrieved by ES, and 20 irrelevant documents were considered for the graph-based representation. Now, as it was mentioned, each graph will contain maximum 100 nodes. A total of 76322 graphs were constructed, where 50% of them correspond to relevant documents and the other 50% to irrelevant ones.

The method was implemented with PyTorch Geometric, and the optimal hyper-parameters were determined via grid search on the validation set. This search included number of layers t, which was searched in {1,2,3,4}, the number of neurons in {16,50,100,200,300} and the batch size in {10,50,80,100}. The results presented in Table 1 were obtained with 3 convolutional layers, 200 neurons and a batch size of 80. Going back to the proposed architecture showed in Figure 1, the GCNN contains, beside its passing layers, an output layer, which predicts a tag equivalent to 1 or 0, indicating relevance or irrelevance for each of the graph representation of query-document interaction, from the testing set. The activation function for the neighbor aggregation is *tanh*, and a *logarithmical softmax* for the output layer. The procedure for the training and testing steps consisted of a division of the graph representation dataset (80% for training and 20% for validation). The selected optimizer is an SGD with a learning rate of 0.01 and a momentum of 0.9. Then, the GCNN was trained for 2000 epochs, and finally the testing set was used to calculate the Mean Average Precision and the Mean Average Recall.

5 RESULTS AND DISCUSSION

In the first place, regarding the results registered in Table 1, the lowest recall the MAP for the re-ranking model reached the highest value and comparing the two solutions the best performance of the system is when the top 5 documents are retrieved and reranked. The reached MAP in this case is equivalent to 0.48. Followed by a MAP of 0.36 for the top 10 documents. These results show that there is a significant improvement between a model that only considers metrics of frequency of occurrence of words, and between one that considers the context of the words in the texts.

On the other hand, Table 1 presents the performance of GBS-DR against the proposed baselines. As it can be seen, GBS-DR outperforms all the document retrieval baselines on the Mean Average Precision. This provides empirical evidence that DR in specialized domain can be improved considering strategies that don't present contextual limitations, which is the case of graph-based strategies. By virtue of the graph structures advantages and the learning process through GCNN, GBS-DR demonstrates a significant improve in the DR tasks. Furthermore, considering the Mean Average Recall the KTRel presents a better performance. The fact that KTRel combines self-attention mechanisms and knowledge-based models, the system can tackle the vocabulary mismatch, of having awareness of contextual information and to focus on the relevant knowledge

Table 1: Performance comparison between different models on the NFCorpus dataset

Algorithm	Mean Average Precision	Mean Average Recall	Mean Average F1 Score
GBS-DR (proposed model)	0.705	0.744	0.724
KTReL [8]	0.241	0.852	0.376
LRDIR [1]	0.162	-	-
BM25+ CEMB	0.326	0.096	0.148
BM25	0.252	0.088	0.130

to improve the search or retrieval of texts. Attention Models were created to address the problem of ¨long-term dependencies¨ [11]. These models have the capacity to extract features from texts and to focus on the relevant terms. Besides, attention mechanisms allow to analyze whole sentences, to make connections between words and its relevant context, so its greater advantage lies in the difference with small-memory and proximity focused networks [12]. On that account, through attention mechanisms the vector space representation for texts turns out to be much more accurate, the caption of contextual information is made even from distant parts of a sentence, so the encoding of information is much more telling when using this approach. This could be the reason why KTRel presents a better recall than the proposed model, besides, GBS-DR uses Word2Vec for the vectorization of words, so in this initial stage the model has the contextual limitation issue.

6 CONCLUSIONS AND FUTURE WORK

In this paper, the GBS-DR was proposed: a similarity graph-based representation for document retrieval in the biomedical domain, which uses the advantages of graph structures to address some of the issues presented in document retrieval. By virtue of the graph structures advantages and the learning process through GCNN, GBS-DR demonstrates a significant improve in the DR tasks. It was empirically demonstrated that adding strategies with no contextual limitations to a neural learning to-rank model (GCNN) sharpens the performance of DR systems. It was also demonstrated that BM25 is strongly improved when integrating vector space methods, like CEMB. Besides, it was exposed that integrating language models based on attention mechanisms to encode context in the vector representation of words, clearly shows an improvement in the results obtained in DR tasks. As future work, it would be desirable to include further approaches to improve the performance and the reliability of the model. To start with the information contained in the queries, a lemmatizing/ POST Tagging process would be useful to mark up the keywords in the query. In this way, each keyword would have a specific token which will allow the understanding of the meaning of the sentence, and then the learning process of the GCNN would be even more precise. During the graph construction stage, the selection of the negative samples could be done more informatively, for instance considering the lowest values of cosine similarity for the selection of the negative samples, could grant a steadier learning process. Finally, it would be appropriate to integrate attention mechanisms for the vectorization of words and concept/entity expansion to increase the contextual information offered to the model.

REFERENCES

[1] V.Boteva, D.Gholipour, A.Sokolov, and S.Riezler. “Full-Text Learning to Rank Dataset for Medical Information Retrieval”(2016)

[2] S.Zhao, C.Su, A.Sboner and F.Wang. “GRAPHENE: A Precise Biomedical Literature Retrieval Engine with Graph Augmented Deep Learning and External Knowledge Empowerment” (2019)

[3] G.Brokos, P.Malakasiotis and I.Androutsopoulos., “Using Centroids of Word Embeddings and Word Mover’s Distance for Biomedical Document Retrieval in Question Answering” (2016).

[4] T.Zhang, B.Liu, D.Niu, K.Lai and Y.Xu. “Multiresolution Graph Attention Networks for Relevance Matching” (2018)

[5] T. Mikolov, K. Chen, G. Corrado and J. Dean, EfficientEstimationofWordRepresentationsinVectorSpace” (2013) International Conference on Learning Representations (2013)

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems* (p./pp. 5998–6008)

[7] X.Yu,W.Xu, Z.Cui and S.Wu1,L.Wang “Graph-based Hierarchical Relevance Matching Signals for Ad-hoc Retrieval” (2021)

[8] J.Frej, J.Chevallet, D.Schwab, “Knowledge Based Transformer Model for Information Retrieval” Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2020), (2020), Samatan, France. ffhal-03263784f

[9] M. Zuckerman and M.Last “Using Graphs for Word Embedding with En-hanced Semantic Relations” (2019)

[10] K.M. Svore and C. J. C. Burges “A Machine Learning Approach for Improved BM25 Retrieval” (2009)

[11] T.Tan “Evolution of Language Models: N-Grams, Word Embeddings, Attention & Transformers” (2020)

[12] C.Nicholson. “A Beginner’s Guide to Attention Mechanisms and Memory Networks” (2019)