SandiDoc at CLEF 2020 - Consumer Health Search : AdHoc IR Task

Sandaru Seneviratne[1], Eleni Daskalaki[1][0000−0002−7665−7039], Md Zakir Hossain[1][0000−0003−1892−831X], and Artem Lenskiy[1][0000−0002−4745−6756]

Research School of Computer Science, College of Engineering and Computer Science, The Australian National University

Abstract. Information retrieval (IR) processes deal with the retrieval of ranked documents based on the similarity among documents and the key words specified in user's query. CLEF 2020 eHealth task on consumer health search adhoc IR addresses the need of improved information retrieval techniques in the health domain to provide relevant information to users. This paper presents our work in CLEF eHealth 2020 consumer health search, on term frequency-inverse document frequency and word vector representation-based techniques adopted in the adhoc IR task. The goal of our work is to experiment on different techniques for information retrieval and look into how different word vector representations of text can affect the final results.

Keywords: Information Retrieval · TF-IDF Score · Word Vector Representations.

1 Introduction

With the increasing expansion of the online content, there has been a growth in online health information retrieval efforts in order to obtain medical knowledge. These efforts, pursued not only by medical specialists but also from the general public, have led to improved mechanisms of health information retrieval. Given the enormous amount of available information, it is vital to provide users with documents fitting to their requests. Information retrieval (IR) can be described as the automatic retrieval of a list of ranked documents that are relevant to a given user query based on similarity measures between the query and the documents. Different theoretical models like boolean, probabilistic, and vector models are used in IR which utilise distinct matching and ranking algorithms to retrieve the documents relevant to a certain query [7].

Most of the early IR systems were based on boolean models [11] which use boolean logic and set theory to represent the presence or the absence of a term in a document respectively. Another major approach for IR is probabilistic retrieval models [11] which make use of the probability of relevance of queries to

documents by calculating the term weights in queries and documents. In vector space models [11], all queries and documents are represented by vectors of an n dimensional vector space where n refers to the number of distinct terms in the collection.

CLEF eHealth 2020 [4] task 2 on consumer health search [3] consists of two sub tasks; adhoc IR and spoken query retrieval. Adhoc IR is a traditional IR task to produce relevant documents to the written queries whereas the spoken query retrieval task utilises spoken queries for the document retrieval. We participated in the sub task 1 of the consumer health search to experiment on adhoc IR.

This paper is organized as follows. Section 2 introduces the data set, queries and other additional resources used in the task. Section 3 describes the methodology used in the experimental setup. In section 4, we present the results and sections 5 and 6 include the discussion and future work respectively.

2 Resources

2.1 Dataset

The document collection used in the document retrieval task was acquired by the common crawl dump of 2018-19. This included web pages of the formats such as HTML, XHTML, XML. The data set used for the task is clefehealth2018_B which is a subset of the initial dataset of 1903 domains. clefehealth_B dataset contains web pages from 1653 website domains and was created by removing a number of websites that were not strictly related to health. The size of this corpus was 294 GB out of which a subset of size 30 GB was used in the task.

2.2 Queries

For the Sub task 1, adhoc IR, 50 topics/queries were provided. These queries were chosen from a set of sample queries collected over 6 months by domain experts. These 50 queries were raw queries with no preprocessing performed beforehand. Fig. 1 provides an example query input which contains the id and the query. The first 3 digits in the id refer to the topic whereas the last 3 digits are used to identify the creator of the query.

```
<query>
        <id> 151001 </id>
        <en> anemia diet therapy </en>
</query>
```

Fig. 1: Example of a query.

2.3 Word Embedding Model

As additional resources, Medical Continuous Bag of Words (CBOW) and Skip-gram word embeddings created using TREC (The Text REtrieval Conference) Medical Records collection were provided. These models use neural network architecture to develop the word representations. In the CBOW architecture, the model takes the context words into account when predicting the target word whereas the Skip-gram model architecture uses the target word to predict the context words [5].

3 Methodology

In this section, we describe the different techniques we used in the document retrieval task. Fig. 2 gives an overview of the complete process which includes preprocessing, representation of queries and documents, and, finally, a matching and ranking algorithm to get the most relevant document list for the queries.

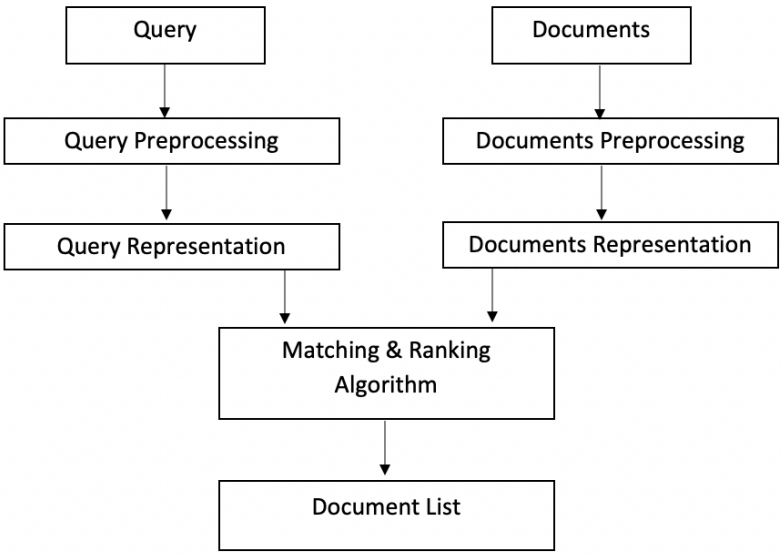

Fig. 2: Elements of a document retrieval system.

3.1 Preprocessing

Preprocessing is an important initial step in natural language processing (NLP) [8] tasks to convert text into a more simplified and an understandable format

so that the NLP and Machine Learning (ML) techniques can perform better. Both the clefehealth_B dataset and queries are raw data with no preprocessing performed on them. In order to obtain clean text from the data set and the queries, we follow different preprocessing steps [9].

Clefehealth_B dataset contains web pages of different formats crawled from the web. These files include the content of the web page along with HTML tags, scripting, and styling. In order to obtain the important content from the web pages, a proper parsing of the web pages is performed using the beautiful-soup library [12]. Converting text to lower case is one of the simplest forms of preprocessing which is useful in entity normalisation. If ignored, this can lead to identifying the same entity as distinct entities which can eventually affect the final result of a system. Both the queries and the text obtained from HTML parsing were converted to lower-case. Next, the digits or the numbers in the text were converted to text in order to facilitate entity normalization. Stop words carry little to no important information in text. Hence, as a next step, stop words were removed in both the queries and the documents using the stop word list provided by nltk library. The punctuation and other unnecessary characters were removed in order to obtain clean text. To ensure that queries and documents are free of spelling errors, spell correction was done using the edit distance of the words and the words in the given word embedding model. Once that was completed, stemming was performed using Porter Stemmer's algorithm to bring each word to its stem word to ensure that different forms of words are identified as one word [10]. Fig. 3 gives an overview of the preprocessing function.

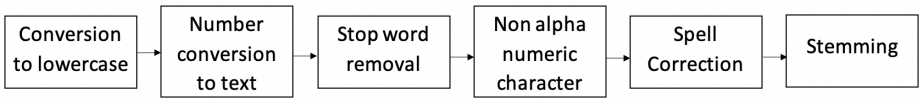

Fig. 3: Elements in the preprocessing function.

3.2 Document Retrieval

This section describes TF-IDF and word embedding based techniques used for the IR task.

TF-IDF score based: TF-IDF (Eq. 1) is a popular IR technique used in many applications [2]. It is a weight (statistical) measure used to evaluate the importance of a word in a document with respect to the whole collection of documents [6]. Number of occurrences of a term in a document (term frequency - TF) and inverse document frequency (IDF) are used to calculate TF-IDF weight. There are different variants of TF and IDF scores used in calculating the relevance of a document to a user query. In our work, we use the variation in equation 1.

Using the TF alone to calculate the scores may give more weight to non-relevant terms. In order to dampen the effect of TF, IDF score is incorporated. However, a linear IDF function may boost the document scores with high IDF terms. To address this issue and dampen the effect of a linear IDF function, log value (sub-linear function) of IDF is considered.

$$w_{i,d} = tf_{i,d} \cdot log(n/df_i) \tag{1}$$

Word Embedding based: Word embeddings can capture semantic meanings among words which is a huge advantage in IR tasks. Word embeddings use distributional hypothesis which focuses on the context of words to derive the word representation[1]. The word embedding model architecture used in the task is Skip-gram which predicts the context words in a window given the target word. Out of the different Skip-gram models, the model with 500 dimensional embedding space and 5 dimensional context window are used in this task.

To represent the documents in the collection using word embedding, we use average, minimum and maximum vector representations using the 100 most frequent terms in the document. Along with the average vector representation obtained (Eq. 2), we average the minimum and maximum vectors to obtain another vector representation (Eq. 3) for the document. Similarly, we obtain two vector representations (average vector and the average of minimum and maximum vectors) for each given query.

$$vector\_representation1_i = \frac{\sum_{n=1}^{100} x_i}{100} \tag{2}$$

$$vector\_representation2_i = \frac{min\_vec_i + max\_vec_i}{2} \tag{3}$$

We calculate the similarity for TF-IDF technique by summing the TF-IDF scores of query tokens in each document and ranking the scores of documents in descending order. For the two vector representations (average vectors, average of the minimum and maximum vectors) we calculate the similarity using cosine similarity to rank the documents with the highest scores in descending order. For each query, we retrieve the 1000 most similar documents as results. Table 1 gives the fields in each row of the results file.

Table 1: Fields in the results file.

Field	Details
qid	query id
Q0	literal Q0
docno	document id
rank	rank of the document
score	similarity score for the document with respect to the query
tag	system identifier

4 Results

The experiment was performed on a subset of size 30 GB of the Clefehealth_B dataset and only the results from the TF-IDF document retrieval algorithm were submitted for evaluation due to time constraints and computational limitations. Using a subset of the dataset has a huge impact on the accuracy of the results since only part of the relevant documents are retrieved, missing a significant number of other relevant documents in the dataset. Table 2 provides the result scores for the IR task using TF-IDF technique for the dataset of 30/294 GB.

Table 2: Results of the adhoc IR task.

Evaluation Metric	Result
Mean Average Precision (MAP)	0.0239
Precision at 10 (P@10)	0.426
Normalized Discounted Cumulative Gain through position 10 (NDCG@10)	0.3235
Accuracy of credibility	0.1744
Relevance-ranked biased precision (RBP_0.95)	0.2981 +0.2934
Credibility-ranked biased precision (cRBP_0.95)	0.1801 +0.2934
Understandability-ranked biased precision (uRBP_0.95)	0.1633 +0.2934

5 Discussion

In this paper, we present our methodology for the adhoc IR subtask of CLEF2020 using TF-IDF score and word vector representations. TF-IDF is considered a simple yet effective algorithm which provides an ideal baseline for IR tasks on which we can develop and expand to more complicated IR algorithms. Despite these advantages, TF-IDF lacks the use of context information compared to other models like word embedding models which take context information into account in developing the embeddings for words. If a user query contains “diabetes” as a key word, TF-IDF algorithm would not consider documents which contain the variation “diabetic” in the IR task. Similarly, the algorithm would not consider documents which contain “diabetec” (misspelled terms) despite how relevant

they are to the user query. In order to produce the most relevant documents using TF-IDF algorithm, it is vital to preprocess the data prior to applying the algorithm which can have a significant effect on the results.

Word embedding models have been successfully used in many NLP and ML tasks since they consider contextual information in developing the representations for words. However, one of the major limitations in word embedding models is that they are unable to identify words similar in text but with different meanings (homonyms) creating a single vector representation for those words. This limitation can be avoided by using approaches which produce multi sense embeddings for words.

6 Future Work

In future, we will further improve and expand algorithms for IR building on the baseline models TF-IDF and word embedding. Moreover, we will expand our current work to incorporate query expansion which can be used to obtain different forms of the original query to improve the results of the IR task. One of the popular query expansion techniques is synonym identification and substitution which is done mostly using existing vocabularies. In the medical domain, vocabularies like UMLS (Unified Medical Language System), SNOMED CT (SNOMED Clinical Terms), OAC-CHV (open-access and collaborative consumer health vocabulary) can be used for query expansion along with word embedding techniques. In addition, we will explore techniques for multi sense embeddings to improve on the word embedding based model for IR.

Acknowledgements

This research was funded by and has been delivered in partnership with Our Health in Our Hands (OHIOH), a strategic initiative of the Australian National University, which aims to transform health care by developing new personalized health technologies and solutions in collaboration with patients, clinicians and health-care providers.

References

1. Croft, W.B., Zamani, H.: Relevance-based Word Embedding. SIGIR '17: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (2017)
2. Fautsch, C., Savoy, J.: Adapting the tf idf Vector-Space Model to Domain Specific Information Retrieval. SAC '10: Proceedings of the 2010 ACM Symposium on Applied Computing (2010), http://www.lucene.apache.org/
3. Goeuriot, L., Suominen, H., Kelly, L., Liu, Z., Pasi, G., Gonzales, G.S., Viviani, M., Xu, C.: Overview of the CLEF eHealth 2020 task 2: Consumer health search with ad hoc and spoken queries. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2020)

4. Goeuriot, L., Suominen, H., Kelly, L., Miranda-Escalada, A., Krallinger, M., Liu, Z., Pasi, G., Saez Gonzales, G., Viviani, M., Xu, C.: Overview of the CLEF eHealth evaluation lab 2020. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., andNicola Ferro, L.C. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020) . LNCS Volume number: 12260, Springer, Heidelberg, Germany (2020)
5. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. Conference: Proceedings of the International Conference on Learning Representations (ICLR 2013) (2013), http://ronan.collobert.com/senna/
6. Ramos, J.: Using TF-IDF to Determine Word Relevance in Document Queries. Tech. rep. (2003)
7. Singhal Google, A.: Modern Information Retrieval: A Brief Overview. Tech. rep. (2001), http://trec.nist.gov
8. Sun, X., Liu, X., Hu, J., Zhu, J.: Empirical Studies on the NLP Techniques for Source Code Data Preprocessing. EAST 2014: Proceedings of the 2014 3rd International Workshop on Evidential Assessment of Software Technologies, vol. 14 (2014). https://doi.org/10.1145/2627508.2627514, http://dx.doi.org/10.1145/2627508.2627514
9. Uysal, A.K., Gunal, S.: The impact of preprocessing on text classification. Information Processing and Management **50**(1), 104–112 (2014). https://doi.org/10.1016/j.ipm.2013.08.006, http://dx.doi.org/10.1016/j.ipm.2013.08.006
10. Willett, P.: The Porter stemming algorithm: Then and now. Electronic library and information systems **40**(3), 219–223 (2006). https://doi.org/10.1108/00330330610681295
11. Yu, B.: Research on information retrieval model based on ontology. EURASIP Journal on Wireless Communications and Networking (2019). https://doi.org/10.1186/s13638-019-1354-z, https://doi.org/10.1186/s13638-019-1354-z
12. Zheng, C., He, G., Peng, Z.: A Study of Web Information Extraction Technology Based on Beautiful Soup. Journal of Computers **10**(6), 381–387 (2015). https://doi.org/10.17706/jcp.10.6.381-387