AUEB at TREC 2015: Clinical Decision Support Track

Giannis Nikolentzos[1,2], Polykarpos Meladianos[1,3], Nektarios Liakis[1], and Michalis Vazirgiannis[1,3]

[1] Athens University of Economics and Business, Greece
[2] Institute for the Management of Information Systems RC "Athena", Greece
[3] LIX, Ecole Polytechnique, France
{nikolentzos,pmeladianos,mvazirg}@aueb.gr, p3100235@dias.aueb.gr

Abstract. One of the goals of clinical decision support systems is to provide physicians information about how to best care for their patients. The Clinical Decision Support track organized by TREC, focuses on developing new techniques to retrieve articles from the biomedical literature relevant to the medical records of the patients. Due to the large volume of the existing literature and the diversity in the biomedical field, this is a very challenging task. This paper describes the two medical information retrieval systems designed by the Athens University of Economics and Business for participation in the 2015 Clinical Decision Support track. The two systems share many common features. Both made use of bigrams along with unigrams for repesenting the documents. Both systems performed automatic query expansion using popular medical knowledge bases. However, the two systems employed different strategies to index the corpus which led to different retrieval methods. One utilized the vector space model with $tf - idf$ term weighting, while the other the vector space model with $tw - idf$ term weighting. The results showed that $tf - idf$ outperformed $tw - idf$.

1 Introduction

Clinical decision making is a very complex process that requires high levels of knowledge and skills. In order to make decisions that lead to high-quality and patient-centered care, physicians often need to obtain adequate information from multiple sources. Biomedical literature forms a valuable source of information for making clinical decisions and physicians often need to consult it for the latest information in patient care. However, due to its immense size, searching for the most relevant and timely information for a particular clinical need is not possible in practice.

Clinical decision support systems are designed to assist physicians with clinical decision making tasks. One of the goals of these systems is to help physicians find information relevant to the medical cases they encounter. As a step towards this direction, the Text REtrieval Conference (TREC) initiated in 2014 the Clinical Decision Support (CDS) track which simulates the requirements of such systems in an attempt to bring new advances to the field with the development of systems able to provide high-quality information. More specifically, each topic within TREC CDS is a medical case narrative that represents an actual medical record. For each topic, there is available a complete description of the patient's case as well as a simplified version that contains less irrelevant information. In addition, the topics are annotated according to the three most common generic clinical question types: diagnosis, test and treatment. The corpus for both the 2014 and the 2015 task was a snapshot of the Open Access Subset[4] of PubMed Central[5] (PMC) consisting of $733,138$ articles in the biomedical domain. For each case report, participants were asked to retrieve the most useful articles for answering a generic clinical question belonging to one of the three types listed above. An example of a case-based topic that was used in the 2015 task is shown in Table 1. In the 2015 CDS track, two rounds of evaluation were conducted (Tasks A & B). In Task A, the structure of all topics was exactly the one described above. In Task B, an additional field was added to the treatment and test topics providing the patient's diagnosis.

No.	Type	Medical Case Narrative
7	diagnosis	**Description**: A 20 yo female college student with no significant past medical history presents with a chief complaint of fatigue. She reports increased sleep and appetite over the past few months as well as difficulty concentrating on her schoolwork. She no longer enjoys spending time with her friends and feels guilty for not spending more time with her family. Her physical exam and laboratory tests, including hemoglobin, hematocrit and thyroid stimulating hormone, are within normal limits. **Summary**: A 22 year old female presents with changes in appetite and sleeping, fatigue, diminished ability to think or concentrate, anhedonia and feelings of guilt.

Table 1. Example of case-based topic used in the 2015 TREC CDS track

In this paper, we present the search models and indexing strategies that we employed for retrieving scientific articles relevant to the topics published by the organizing committee of CDS. We chose to incorporate both unigrams and bigrams in our document representation. Hence, besides unigrams, bigrams were also indexed. As regards the weighting of the terms within each document, we employed the well-known tf score and we also introduced a new score, denoted by tw, which corresponds to the importance of the terms within the graph-of-words document representation. We explored two different retrieval methods: the vector space model with $tf-idf$ similarity and the vector space model with $tw-idf$ similarity. Finally, we utilized popular external biomedical knowledge resources (MetaMap, Wikipedia) for automatic query expansion.

[4] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[5] http://www.ncbi.nlm.nih.gov/pmc/

The rest of this paper is organized as follows. Section 2 provides a detailed description of our proposed approaches. Section 3 presents the evaluation results of these approaches. Finally, Section 4 summarizes the work and presents potential future work.

2 Methodology

In this section, we present the various indexing and retrieval strategies that we employed as well as the preprocessing phase that preceded indexing. For both indexing and retrieval, we used Lucene[6], a well-known search engine in Information Retrieval.

2.1 Preprocessing Phase

We obtained the collection of documents from the official site of the task[7]. We chose to index only the text of the articles and to ignore the supplemental material. The full text of each of the $733,138$ articles is represented as an *NXML* file (*XML* encoded using the NLM Journal Archiving and Interchange Tag Library). For each article, we extracted its plain text by removing any tags and irrelevant information. The next step was to perform standard text processing tasks such as tokenization, stopword, punctuation and special character removal, and stemming. The stemming was performed using Porter's algorithm [8].

2.2 Indexing Strategy

After the preprocessing phase was completed, the processed text of all articles was transformed into an inverted index in order to make term-based search more efficient.

One of the main features of our system is the use of bigrams. An n-gram is a contiguous sequence of n items from a given sequence of text. In our case, the items correspond to terms. More specifically, we indexed both the terms appearing in the text and the bigrams constructed by these terms. Due to the highly technical terminology of the biomedical literature, pairs of terms may appear often together in the same order. Hence, although the use of bigrams would make the indexing and retrieval phases slower, it could prove beneficial for the retrieval of relevant articles.

We created two different indexes. The first index used Lucene's default scoring method: a vector space model using $tf-idf$ weighting and cosine similarity. The second index was created by replacing the $tf-idf$ scores of the terms with their $tw-idf$ scores, where tw is a score measuring the importance of each term in the graph-of-words representation of documents described below.

Besides the traditional vector space model, we also chose to represent each document as a statistical graph-of-words, following earlier approaches in keyword

[6] https://lucene.apache.org/

[7] http://www.trec-cds.org/2015.html#documents

extraction [7,6] and more recent ones in ad hoc IR [3,9] and in summarization [5].

After the preprocessing and bigram generation phases, each document is transformed into an unweighted, undirected graph whose vertices represent unique terms and whose edges represent co-occurrences of the connected terms within a fixed-size window. Hence, besides the terms (vertices), the graph-of-words representation of text also models the relationships between them (edges). All the words present in a document have some relationships with one another, modulo a window size outside of which the relationship is not taken into consideration, and graphs are able to capture these dependencies.

We next give in Figure 1 an example of the graph-of-words representation

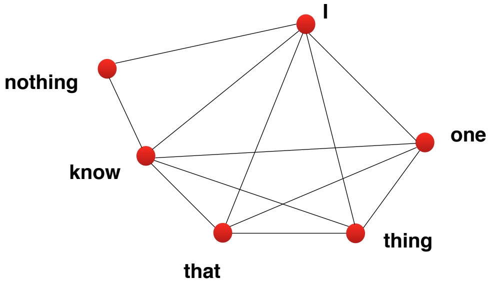

Fig. 1. Example of the graph representation of a textual document. Vertices correspond to words and edges indicate co-occurrence of the connected words in a window of size 3 in the text.

of Socrates's well-known saying: "I know one thing: that I know nothing". For illustration purposes, no bigrams are constructed. In addition, only the colon is removed and no other text processing tasks are performed. Each vertex represents a unique term and each edge a co-occurrence of the two terms in at least one window of size 3, i.e. it captures trigram relationships. Hence, each word (vertex) is connected with an edge with each one of its two preceding and two following words, if any. As regards the documents belonging to the CDS track's corpus, we set the size of the window equal to 6 as the number of terms in the documents is generally large.

In order to measure the importance of each vertex (word) within the graph-of-words, we employed PageRank [4], a well-known ranking algorithm. Let $G = (V, E)$ be a simple undirected graph where V is the set of vertices and E the set of edges. We will denote $ne(v)$ the neighbors of vertex $v \in V$, that is, the vertices that are adjacent to v. The number of neighbors of vertex $v \in V$ is called the degree of v, and we denote it by $deg(v)$. The score of a vertex v is defined as follows:

$$S(v) = (1 - d) + d \sum_{v' \in ne(v)} \frac{1}{deg(v')} S(v') \tag{1}$$

where d is a damping factor that integrates into the model the probability of jumping from a given vertex to another random vertex in the graph. The damping factor d can take values between 0 and 1 and it is usually set to 0.85. For our experiments, we also set d equal to this value.

After computing the PageRank scores for all the vertices of the graph, we determine the tw score of each term using the following formula:

$$tw = \sqrt{S(u)} \tag{2}$$

where tw is the score of the term corresponding to vertex u in the graph.

2.3 Retrieval Strategy

At the retrieval stage, the documents are ranked in decreasing order by relevance criteria with respect to a query. Although the descriptions of the medical cases are generally long, we performed query expansion in order to produce more useful queries that are likely to retrieve more relevant documents.

Query expansion is the process of adding additional terms to a query in order to improve the retrieval performance. It has proven to be an effective strategy for improving search results in biomedical Information Retrieval [10, 2]. Note that most of the participants in the 2014 CDS task performed query expansion by adding extra terms to the original topics. In the biomedical domain, different terms may correspond to the same concept. A document relevant to a medical case narrative may not be retrieved if it makes exclusive use of a term different from the one used in the narrative to describe the same concept. To deal with this problem, we included a query expansion module in our system. This module expands the queries by adding the synonyms of the clinical terms that appear in them. The module uses MetaMap [1], a popular tool for recognizing clinical concepts in the text.

In Task B, besides MetaMap, we also performed query expansion using information extracted from Wikipedia[8], a common source of knowledge for natural language processing tasks. In this task, the organizing committee provided participants with a diagnosis field for the treatment and test topics. Given a patient's diagnosis, the approach that we took was to retrieve the corresponding to the diagnosis Wikipedia article and then to expand the query by appending the summary of the Wikipedia article to it. Note here that for 4 out of the 20 topics containing a diagnosis field, no relevant Wikipedia articles could be retrieved. Hence, query expansion was performed only for the remaining 16 topics.

3 Evaluation

In this section, we provide results regarding our participation in the 2015 CDS track. As mentioned earlier, in the 2015 task, two rounds of evaluation were conducted (Tasks A & B). The track received a total of 178 runs from 36 different

[8] www.wikipedia.org

groups. For each topic, the following four evaluation metrics were computed: inferred average precision (infAP), inferred normalized discounted cumulative gain (infNDCG), precision at R (R-prec) and precision at 10 (P@10). More information about these metrics is given in [11]. Four automatic runs were submitted for evaluation, two in Task A and two in Task B. In Task A, the following two runs were submitted:

1. **tw_bi_ex**: uses a vector space model including both unigrams and bigrams, $tw - idf$ weighting and query expansion using MetaMap.
2. **tf_bi_ex**: uses a vector space model including both unigrams and bigrams, $tf - idf$ weighting and query expansion using MetaMap.

In Task B, the **tw_bi_ex** run was exactly the same as in Task A, while in the case of the **tf_bi_ex** run, the queries were further expanded using information from Wikipedia. For each run, we calculated the average of each metric over all topics. We also calculated the average of each metric over topics belonging to each clinical question type. The results for the two tasks of the 2015 CDS track are shown in Table 2. The upper table corresponds to **tw_bi_ex**, while the lower corresponds to **tf_bi_ex**. From the table, we can observe that **tf_bi_ex** had the best overall results for each metric on both tasks, while **tw_bi_ex** did not perform as well. Hence, in contrast to other tasks [3], in retrieving biomedical articles relevant to medical cases, $tw - idf$ weighting failed to outperform $tf - idf$ weighting. As regards the performance per clinical question type, in Task A, **tf_bi_ex** produced better infNDCG, R-prec and P@10 scores than **tw_bi_ex** on all topics. However, **tw_bi_ex** managed to outperform **tf_bi_ex**, in terms of infAP score, on diagnosis and test topics. In Task B, **tf_bi_ex** dominated on

	Task A				**Task B**			
	infAP	**infNDCG**	**R-prec**	**P@10**	**infAP**	**infNDCG**	**R-prec**	**P@10**
diagnosis	0.0254	0.1546	0.1153	0.2900	0.0246	0.1525	0.1122	0.2800
test	0.0270	0.1940	0.1498	0.3200	0.0258	0.1918	0.1483	0.3200
treatment	0.0675	0.2377	0.1889	0.4800	0.0674	0.2376	0.1886	0.4800
overall	0.0400	0.1954	0.1513	0.3633	0.0393	0.1939	0.1497	0.3200

	Task A				**Task B**			
	infAP	**infNDCG**	**R-prec**	**P@10**	**infAP**	**infNDCG**	**R-prec**	**P@10**
diagnosis	0.0250	0.1698	0.1333	0.3000	0.0250	0.1698	0.1333	0.3000
test	0.0263	0.2050	0.1662	0.3200	0.0669	0.3599	0.2723	0.6100
treatment	0.0782	0.2602	0.2156	0.5100	0.1050	0.3396	0.2867	0.5100
overall	0.0432	0.2117	0.1717	0.3767	0.0656	0.2898	0.2308	0.4733

Table 2. Retrieval scores of **tw_bi_ex** (upper table) and **tf_bi_ex** (lower table) on the 2015 CDS Tasks A & B.

all topics and for all metrics. In Task A, both **tw_bi_ex** and **tf_bi_ex** are most effective in retrieving relevant documents for topics belonging to the treatment category and next to the test category. Conversely, in Task B, using the extra information provided by the diagnosis field, **tf_bi_ex** managed to retrieve more relevant documents for topics belonging to the test category than for topics belonging to the treatment category.

To provide a comparative analysis of the performance of our runs with the runs submitted by the other participants, we present the same information for the average of the **maximum** and the average of the **median** scores. More specifically, Table 3 illustrates the average of the **maximum** and the average of the **median** scores of each metric over all topics and over topics belonging to each clinical question type for both tasks of the 2015 CDS track. The upper table corresponds to **maximum**, while the lower corresponds to **median**. On average, the scores produced by our runs were close to the **median** scores, while they were way below the **maximum** scores. In Task A, **tw_bi_ex** performed below the **median** for all metrics except for P@10, while **tf_bi_ex** performed above the **median** for all metrics. In TaskB, **tw_bi_ex** performed below the **median** for all metrics. However, this is not a fair comparison. As mentioned earlier, **tw_bi_ex** does not exploit the information provided by the diagnosis field. On the other hand, **tf_bi_ex** again managed to performed above the **median** for all metrics.

	Task A				**Task B**			
	infAP	**infNDCG**	**R-prec**	**P@10**	**infAP**	**infNDCG**	**R-prec**	**P@10**
diagnosis	0.0900	0.4011	0.2809	0.6500	0.0916	0.4334	0.2877	0.6800
test	0.0845	0.4042	0.2768	0.7000	0.1377	0.5502	0.3872	0.8300
treatment	0.2030	0.5144	0.4089	0.7000	0.2718	0.6208	0.5043	0.8400
overall	0.1258	0.4399	0.3222	0.6833	0.1670	0.5348	0.3939	0.7833

	Task A				**Task B**			
	infAP	**infNDCG**	**R-prec**	**P@10**	**infAP**	**infNDCG**	**R-prec**	**P@10**
diagnosis	0.0244	0.1784	0.1373	0.3100	0.0267	0.1872	0.1389	0.3000
test	0.0236	0.1781	0.1362	0.3100	0.0461	0.2920	0.2075	0.5100
treatment	0.0762	0.2551	0.2111	0.4100	0.1171	0.3589	0.2905	0.5400
overall	0.0414	0.2038	0.1615	0.3433	0.0633	0.2794	0.2123	0.4500

Table 3. Maximum (upper table) and **median** (lower table) scores of automatic runs on the 2015 CDS Tasks A & B.

We next investigate how challenging was the retrieval of relevant articles for each one of the 30 topics. Figure 2 illustrates the P@10 scores of **tw_bi_ex**, **tf_bi_ex**, **maximum** and **median** for all 30 topics of Task A. For some topics, the retrieval of relevant articles was a really challenging task. For example, for topics 18, 20, 25, 27 and 28, more than half of the participant systems could not

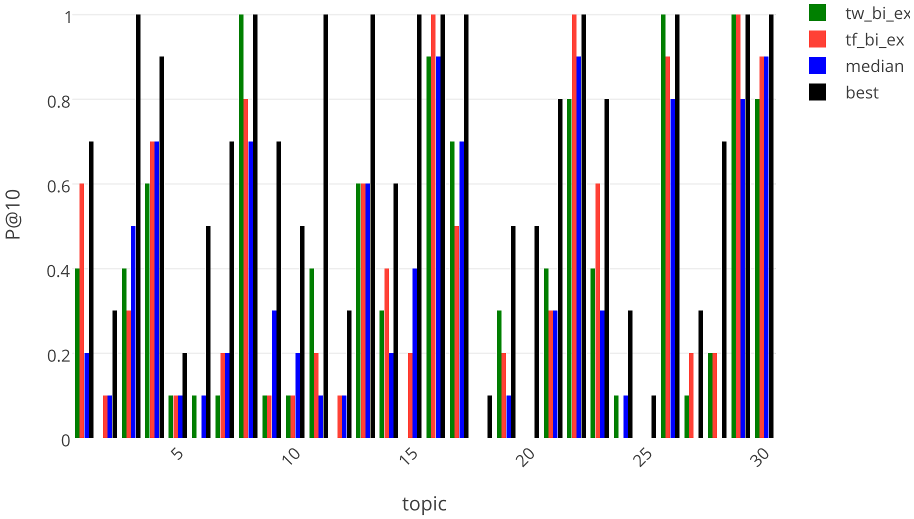

Fig. 2. P@10 scores for all 30 topics of Task A.

find one single relevant document in the top 10. For other topics, such as topics 16, 22 and 30, the results were high in general. The **median** P@10 scores for all these three topics were 0.9000. The same scores for all 30 topics of Task B are shown in Figure 3. The additional information provided by the diagnosis field increased the general performance of the participant systems. More specifically, in Task B, only for two topics (2 and 25), more than half of the participant systems could not find one single relevant document in the top 10. Hence, the number of such topics decreased from 5 in Task A to 2 in Task B. In addition, the number of topics for which the participant systems managed to retrieve several relevant documents in the top 10 increased significantly compared to Task A. For example, the **median** P@10 scores for topics 16, 22, 26, 29 and 30 were 1.0000, 1.0000, 0.9000, 0.9000 and 0.9000 respectively.

4 Conclusion

Improving the performance of systems that retrieve documents relevant to medical case descriptions is an important challenge for Information Retrieval. TREC contributed to this end by organizing the CDS track for second year in a row. In this paper, we presented the different systems that we built for participating in the 2015 CDS track. Our participation focused on the evaluation of various language models, document representations, term weighting models and query

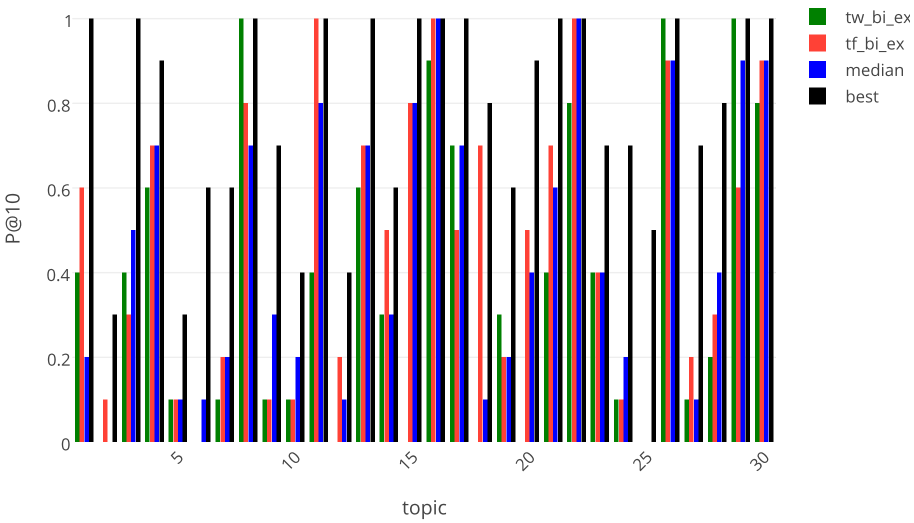

Fig. 3. P@10 scores for all 30 topics of Task B

expansion models. The system that weighted the terms using the popular tf term weighting scheme exhibited better performance than the system that weighted the terms based on their importance within the graph-of-words representation of documents. In general, our proposed methods yielded relatively good results and can serve as a good starting point for future participation in the track. In terms of future directions of research, we would like to investigate how to more properly utilize external biomedical knowledge resources.

References

1. Aronson, A.R.: Effective Mapping of Biomedical Text to the UMLS Metathesaurus: the MetaMap Program. In: Proceedings of the American Medical Informatics Association Symposium. pp. 17–21 (2001)
2. Aronson, A.R., Rindflesch, T.C.: Query Expansion Using the UMLS Metathesaurus. In: Proceedings of the American Medical Informatics Association Symposium. pp. 485–489 (1997)
3. Blanco, R., Lioma, C.: Graph-based term weighting for information retrieval. Information Retrieval 15(1), 54–92 (2012)
4. Brin, S., Page, L.: Reprint of: The anatomy of a large-scale hypertextual web search engine. Computer Networks 56(18), 3825–3833 (2012)
5. Meladianos, P., Nikolentzos, G., Rousseau, F., Stavrakas, Y., Vazirgiannis, M.: Degeneracy-Based Real-Time Sub-Event Detection in Twitter Stream. In: Ninth International AAAI Conference on Web and Social Media. pp. 248–257 (2015)

6. Mihalcea, R., Tarau, P.: TextRank: Bringing Order into Texts. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 404–411 (2004)
7. Ohsawa, Y., Benson, N.E., Yachida, M.: Keygraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In: Proceedings of the Advances in Digital Libraries Conference. pp. 12–18 (1998)
8. Porter, M.F.: An algorithm for suffix stripping. Program 14(3), 130–137 (1980)
9. Rousseau, F., Vazirgiannis, M.: Graph-of-word and TW-IDF: New Approach to Ad Hoc IR. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. pp. 59–68 (2013)
10. Srinivasan, P.: Optimal document-indexing vocabulary for medline. Information Processing & Management 32(5), 503–514 (1996)
11. Yilmaz, E., Kanoulas, E., Aslam, J.A.: A Simple and Efficient Sampling Method for Estimating AP and NDCG. In: Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 603–610 (2008)