TUW @ TREC Clinical Decision Support Track

João Palotti, Navid Rekabsaz, Linda Anderson, and Allan Hanbury

Institute of Software Technology and Interactive Systems
Vienna University of Technology, Austria
{palotti, rekabsaz, anderson, hanbury}@ifs.tuwien.ac.at

Abstract. This document describes the participation of Vienna University of Technology in the TREC Clinical Decision Support Track 2014. Four different search models are investigated, as well as different strategies to index the corpus and to extract the most relevant information from the topics. Our results conclude that BM25 and Vector Space Model had similar performance for P@10 and inferred NDCG.

Keywords: Medical Information Retrieval, Evaluation

1 Introduction

Searching for health has become a common task nowadays. Pew Research Center estimates that 80% of the American population uses the Web to seek health information [2]. In line with this trend, various health-related campaigns were proposed. Some examples are the TREC Genomics Track [7] which ran from 2003 to 2007, the TREC Medical Records Track [9] running in 2011 and 2012, the ImageCLEFmed Track on medical image retrieval [4,5] running between 2003 and 2013, and the ShARe/CLEF eHealth Evaluation Lab [8,3] running in 2013 and 2014. Here we briefly describe the goals of the first TREC Clinical Decision Support Track (TREC-CDS) and the participation of Vienna University of Technology.

The TREC-CDS is focused on physicians searching for relevant information for patient care. As document collection, it uses the open access subset of PubMed Central (PMC), containing a total of 733,138 articles. The topics are divided into three main types: diagnosis, test and treatment. Figure 1 shows a diagnosis query.

As there was no development set available, we decided to experiment with different search models and indexing possibilities, trying to build a initial foundation for our future participation next year.

Our Contribution

In this paper, we experiment and evaluate a large variety of search models and indexing strategies, as well as ways of combining different models and indexes.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **NOV 2014**	2. REPORT TYPE	3. DATES COVERED **00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE
TUW @ TREC Clinical Decision Support Track

5a. CONTRACT NUMBER
5b. GRANT NUMBER
5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER
5e. TASK NUMBER
5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Vienna University of Technology,Institute of Software Technology and Interactive Systems,Austria,

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
presented in the proceedings of the Twenty-Third Text REtrieval Conference (TREC 2014) held in Gaithersburg, Maryland, November 19-21, 2014. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA).

14. ABSTRACT
This document describes the participation of Vienna University of Technology in the TREC Clinical Decision Support Track 2014. Four di erent search models are investigated, as well as di erent strategies to index the corpus and to extract the most relevant information from the topics. Our results conclude that BM25 and Vector Space Model had similar performance for P@10 and inferred NDCG.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**7**	

```
<topic number="8" type="diagnosis">
   <description>
     A 62-year-old man sees a neurologist for progressive memory loss and
     jerking movements of the lower extremities. Neurologic examination
     confirms severe cognitive deficits and memory dysfunction. An
     electroencephalogram shows generalized periodic sharp waves.
     Neuroimaging studies show moderately advanced cerebral atrophy.
     A cortical biopsy shows diffuse vacuolar changes of the gray matter
     with reactive astrocytosis but no inflammatory infiltration.
   </description>
   <summary>
     62-year-old man with progressive memory loss and involuntary leg
     movements. Brain MRI reveals cortical atrophy, and cortical biopsy
     shows vacuolar gray matter changes with reactive astrocytosis.
   </summary>
</topic>
```

Fig. 1: Example of a diagnosis query

2 Experiments

In our experiments, we explore several different search techniques, IR-system, as well as different indexing strategies. In this section all the different configurations used will be described in details. In Section 2.1, we explain how we create three varieties of index using the MeSH hierarchy. Thereafter, in Section 2.2 we explain our query formulation method, where we make use of Metamap to retain only the most important concepts from each topic. In the Sections 2.3, 2.4 and 2.5 we briefly explain the 3 different IR-systems that we use for our runs: Run1 (Indri), Run2 (Lucene), and Run3 (Xapian). For each system, we generate 6 runs: a combination of the 3 indices methods from Section 2.1 and 2 query strategies from Section 2.2. We merge the scores of each run into a final run for each system. For Run4, we combine the documents from the previous 3 runs, as we explain in Section 2.6. Finally, we explore Word2Vec in our Run5, explained in Section 2.7.

2.1 Indexing Concepts

We take advantage of the Medical Subject Headings (MeSH[1]) hierarchy to keep only the important concepts of each document in the collection. MeSH has an hierarchical structure for a set of terms named descriptors as shown in Figure 2. The hierarchical structure makes it possible to narrow the scope of the terms. It is updated every year and the 2014 version has 27,149 descriptors.

Based on MeSH hierarchy, we create 3 types of indexes:

1. All words: we index the documents as they are, without removing and word;

[1] http://www.nlm.nih.gov/mesh/MBrowser.html

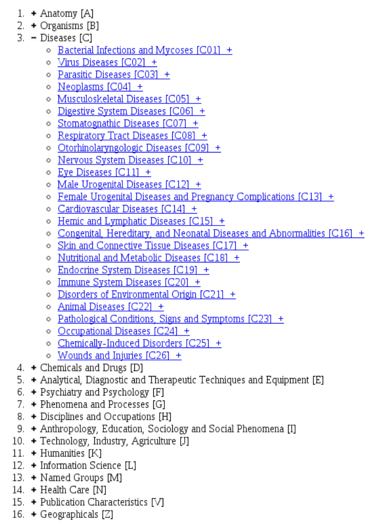

Fig. 2: MeSH hierarchy with the disease branch expanded

2. MeSH vocabulary: we exclude all words in a document that are not present in the MeSH hierarchy;
3. MeSH-CD: we exclude all words in a document that are not present in the branch *(C) - Disease* or *(D) - Chemicals and Drugs* of the MeSH hierarchy.

We use all 3 indexes for runs 1, 2, 3 and 4, and only MeSH-CD for run 5, as describe in Table 1. For all runs, a script to lowercase and remove punctuation is also used.

Runs	**System**		**Indexing Variants**			**Query Variants**	
	Model	**Search Engine**	**All Words**	**MeSH voc.**	**MeSH-CD**	**Whole Desc.**	**Metamap Filter**
Run1	Language Model	Indri	✓	✓	✓	✓	✓
Run2	Vector Space Model	Lucene	✓	✓	✓	✓	✓
Run3	BM25	Xapian	✓	✓	✓	✓	✓
Run4	-	Combo	✓	✓	✓	✓	✓
Run5	-	Word2Vec			✓		✓

Table 1: Summary description of all 5 runs

2.2 Selecting Terms in the Topics

We employ NLM's Metamap (version 2013) with default processing options [1] to annotate all the topics. Metamap maps the topics to UMLS concepts and semantic types. There are a total of 133 semantic types, but some of them (e.g., *Clinical Drug* or *Disease or Syndrome*) are more important than others in our experiments[2]. For example, the last sentence in the description part of Figure 1 is: *"A 62-year-old man sees a neurologist for progressive memory loss and jerking movements of the lower extremities"* from which Metamap identifies concepts such as:

- Concept: */year (per year)* – Semantic Type: *Temporal Concept*
- Concept: *Old* – Semantic type: *Temporal Concept*
- Concept: *MAN (Male gender)* – Semantic type: *Finding*
- Concept: *sees (Vision)* – Semantic type: *Organism Function*
- ...
- Concept: *(Lower - spatial qualifier)*– Semantic type: *Spatial Concept*

In an automatic manner, we kept only the concepts in which the semantic types are related to symptoms, diseases or treatments (based on [6]): *man*, *memory loss*, *jerking movements*, and *lower extremities*.

For each topic, we can:

1. use the description of the topic as the query;
2. use only the keywords related to symptom, diseases or treatments, provided by Metamap semantic types.

For runs 1, 2, 3 and 4 we generated runs both possibilities. For Run5, we only generated runs using only the second possibility.

2.3 Run1

Run1 was based on Indri[3]. Indri is a search engine from the Lemur project, mainly based on Language Modeling as retrieval model. We used only the *#combine* operator in our experiments. Six runs were generated: three different indexing strategies combined with two different ways to formulate the queries. The runs were combined simply adding the scores for each document.

2.4 Run2

Lucene[4] is a text search engine written in Java and supported by the Apache Foundation. The default search model of Lucene is the Vector Space Model (VSM), and it was used with the default parameters. As for Run1, six runs were generated and combined summing the scores of each individual document.

2.5 Run3

Xapian[5] is also an open source search engine. It is written in C++ and has BM25 weighting scheme as its default. The scores of the six run created were also summed for each document.

[2] A complete list of every semantic type can be found here: http://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml

[3] http://www.lemurproject.org/indri/

[4] http://lucene.apache.org/

[5] http://xapian.org/

2.6 Run4

Our Run4 is the combination of all the previous runs. However, instead of using the raw scores provided by the systems, we used the position a document had in each run as its score (1/position).

2.7 Run5

Word2Vec[6] provides vector representation of words by using deep learning. We had to compared each word in the query with each word in the documents, in a quadratic procedure. Therefore, we used only the MeSH-CD indexing strategy and the Metamap strategy for building the queries.

3 Results

We detail the results for all 30 topics in Figure 3. There were some very difficult topics, such as 3, 9, 23 and 25, in which more than 50% of all participant systems could not find one single relevant document in the top 10. For other topics, such topic 4 and 27, the results were in general high. On average, our systems, in special the ones using Xapian and Lucene as base, were as good as the median system for both P@10 and inferred NDCG.

In general, Run1 was our worst run, performing much worse than the others. Run5 also did not perform so well, but it can be explained by the fact that only the smaller indexing strategy (MeSH-CD) and Metamap queries were used for this system. In any case, a detailed investigation of the performance of these two runs need to be carried in the future. Run2 and Run3 were our best runs, Run3 had slight better performance for P@10, but Run2 was better for inferred NDCG. Run4 was stable enough to perform relatively well even after the terrible performance of Run1. Table 2 compares the averaged results for all 5 runs, the median and the best system for each topic.

Runs	P10	InfNDCG	infAP	RPrec
Best	0.71	0.520	0.180	0.350
Median	0.23	0.151	0.032	0.126
Run1	0.02	0.017	0.001	0.007
Run2	0.28	**0.193**	**0.057**	**0.174**
Run3	**0.29**	0.171	0.042	0.152
Run4	0.23	0.152	0.033	0.141
Run5	0.14	0.059	0.009	0.040

Table 2: Results averaged over the 30 topics for each of our 5 runs, the Best and Median system.

[6] https://code.google.com/p/word2vec/

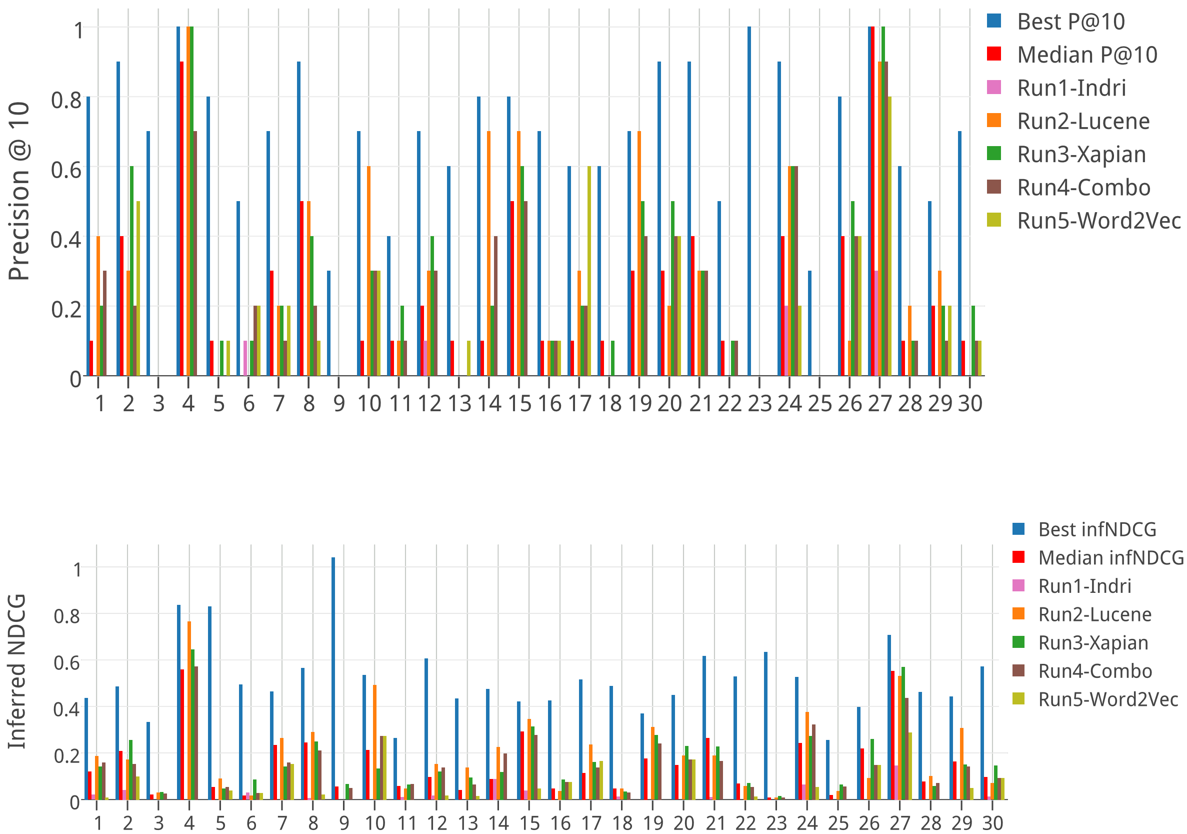

Fig. 3: Precision at 10 and Inferred NDCG for all 30 topics.

4 Conclusion and Future Work

Improving search systems for health related documents is an important challenge for information retrieval researchers. In this work, we focused on creating a robust baseline system, testing different search models and indexing alternatives and possible ensembles.

Our experiments have shown that Lucene, using Vector Space Model, and Xapian, using BM25, had very similar performances. An ensemble of these two can lead for better results, and it is one of our future work. Also, investigating what caused the Language Model of Indri to perform so bad is left as an important future work.

Acknowledgements

This research was funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n°257528 (KHRESMOI) and partly funded by the Austrian Science Fund (FWF) project number I1094-N23 (MUCKE).

References

1. A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. pages 17–21, 2001.
2. Susannah Fox. *Health topics: 80% of internet users look for health information online*. Pew Internet & American Life Project, 2011.
3. Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Tobias Schreck, Gondy Leroy, Danielle L. Mowery, Sumithra Velupillai, Wendy W. Chapman, David Martínez, Guido Zuccon, and João R. M. Palotti. Overview of the share/clef ehealth evaluation lab 2014. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction - 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings*, pages 172–191, 2014.
4. Henning Müller, Paul Clough, Thomas Deselaers, and Barbara Caputo. *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*. Springer Publishing Company, Incorporated, 1st edition, 2010.
5. Henning Müller, Alba García Seco de Herrera, Jayashree Kalpathy-Cramer, Dina Demner-Fushman, Sameer Antani, and Ivan Eggel. Overview of the imageclef 2012 medical image retrieval and classification tasks. In *CLEF 2012 working notes*, 2012.
6. João R. M. Palotti, Veronika Stefanov, and Allan Hanbury. User intent behind medical queries: an evaluation of entity mapping approaches with metamap and freebase. In *IIiX*, pages 283–286, 2014.
7. Phoebe M. Roberts, Aaron M. Cohen, and William R. Hersh. Tasks, topics and relevance judging for the TREC genomics track: five years of experience evaluating biomedical text information retrieval systems. *Inf. Retr.*, 12(1):81–97, 2009.
8. Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy Webber Chapman, Guergana K. Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, Gareth J. F. Jones, Johannes Leveling, Liadh Kelly, Lorraine Goeuriot, David Martínez, and Guido Zuccon. Overview of the share/clef ehealth evaluation lab 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization - 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings*, pages 212–231, 2013.
9. Ellen M. Voorhees. The TREC medical records track. In *ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics. ACM-BCB 2013, Washington, DC, USA*, page 239, 2013.