WSU-IR at TREC 2015 Clinical Decision Support Track: Joint Weighting of Explicit and Latent Medical Query Concepts from Diverse Sources

Saeid Balaneshin-kordan, Alexander Kotov, and Railan Xisto*

Department of Computer Science, Wayne State University, Detroit MI 48226, USA
{saeid,kotov}@wayne.edu,railan.xisto@gmail.com,

Abstract. This paper describes participation of WSU-IR group in TREC 2015 Clinical Decision Support (CDS) track. We present a Markov Random Fields-based retrieval model and an optimization method for jointly weighting statistical and semantic unigram, bigram and multi-phrase concepts from the query and PRF documents as well as three specific instantiations of this model that we used to obtain the runs submitted for each task in this track. These instantiations consider different types of concepts and use different parts of topics as queries.

1 Introduction

An information retrieval scenario, in which a verbose query describes a patient case in several sentences and aims at finding relevant articles in medical literature, is fairly common in clinical practice. Such queries typically correspond to complex information needs, which involve a large number of concepts of different types, including patient demographics, symptoms of a disease or laboratory test results. For example, the query "*A 4-year-old girl presents with persistent fever for the past week. The parents report a spike at 104°F. The parents brought the child to the emergency room when they noticed erythematous rash on the girl's trunk. Physical examination reveals strawberry red tongue, red and cracked lips, and swollen red hands. The whites of both eyes are red with no discharge.*", includes the query concepts that indicate the age and gender of a patient, describes several symptoms, such as *erythematous rash*, and test results, such as *revealed swollen red hands* and *strawberry red tongue*, as well as indicates a possible diagnosis, such as *strawberry tongue* (also known as *Kawasaki disease*). Although such queries are fairly long, only a fraction of concepts corresponding to an information need underlying those queries are directly mentioned in them, such as the concept *strawberry tongue* in the query above (i.e. explicit concepts), while many other concepts representing the same information need do not occur in the queries themselves, but can be found in the pseudo-relevance feedback (PRF) documents (i.e. latent concepts). For example, the concept *Kawasaki disease* that is not explicitly mentioned in the above query, can be found in an article with the PubMed unique identifier (PMID) 3625593 (i.e., [5]), which is among the top

* undergraduate research intern under BSMP program

retrieved documents for this query. Therefore, accurately answering such information needs requires not only capturing all explicit and latent query concepts, but also determining their relative importance.

Previously proposed approaches to identify and weight query concepts are either based only on semantics [9, 16, 6, 17, 22] or are purely statistical [18, 10, 2, 3, 7, 11, 21]. Each of these two types of approaches are able to identify only certain types of concepts. For example, [9] identifies and utilizes only the concepts from the Unified Medical Language System (UMLS) that are extracted using the MetaMap tool [1] from PRF documents. Single-word and multi-word statistical concepts from the query and single-word concepts from PRF documents have been show to be effective for ad-hoc retrieval in [11, 3]. An bag-of-words retrieval model utilizing medical concepts from PRF documents for query expansion was proposed in [16]. Choi et al. [4] proposed a method to represent multi-word UMLS concepts using sequential dependencies between their words.

In this work, we present a Markov Random Fields-based retrieval model and an optimization method for jointly weighting statistical and semantic unigram, bigram and multi-phrase concepts from the query and PRF documents as well as three specific instantiations of this model that we used to obtain the runs submitted for each task in TREC 2015 Clinical Decision Support (CDS) track. These instantiations consider different types of concepts and use different parts of topics as queries.

2 Methods

2.1 Retrieval Model

In this section, we provide the details of the six runs that were submitted to TREC 2015 CDS track. Three of these runs were submitted for Task A and three others were submitted for Task B of this track. The runs submitted for Task B consider the *diagnosis* section provided for some of the topics in this task. These *diagnosis* sections are considered as n-gram concepts and added with the optimized weights to the expanded queries. As mentioned in [15], considering all of the runs in TREC 2014 CDS track, a very small difference in retrieval performance is observed when the query types (i.e., "Diagnosis", "Test", and "Treatment") are taken into account. Therefore, query types are not taken into account in this work.

In this work, we assume that the concepts representing the information need underlying the query exist both in the query itself as well as in other concept sources, such as PRF documents. We also assume the existence of sequential dependencies between the adjacent terms of multi-word concepts, which can be accounted for in retrieval by using the Markov Random Field (MRF) model [10]. In particular, our retrieval model builds upon the Markov Random Field-based Parameterized Query Expansion (PQE) framework [3], which assumes that the information need underlying a multi-term query can be categorized using three query concept types (unigrams, ordered bigrams, and unordered

bigrams), each of which is associated with its own matching function. We extend this framework by considering more fine-grained concept types, depending on whether the concepts of the above three types occur in the query itself (including the multi-word UMLS concepts) or in the PRF documents, and thus providing a more flexible concept matching strategy. Specifically, in our retrieval model, contribution of a query concept c to the retrieval score of document D, in which it occurs, is determined as:

$$sc(c, D) = \sum_{T \in \mathbb{T}} \lambda_T f_T(c, D) \tag{1}$$

where $\mathbb{T}$ is a set of all concept types, to which concept c belongs (a query concept can belong to several concept types; for example, if it occurs in both the query and the PRF documents) and λ_T is the relative importance weight of the concepts of type T (all concepts of the same type are assigned the same weight). The final retrieval score of document D given a query is determined as a linear combination of contributions of all query concepts occurring in D:

$$\begin{aligned} sc(Q, D) &= \sum_{c \in \mathbb{C}} \mathbb{I}_c sc(c, D) \\ &= \sum_{c \in \mathbb{C}} \mathbb{I}_c \sum_{T \in \mathbb{T}} \lambda_T f_T(c, D) \end{aligned} \tag{2}$$

where $\mathbb{C}$ is the set of all explicit and latent query concepts, $\mathbb{I}_c$ is an indicator function that determines whether concept c is considered (i.e. it takes the value of 1) or not (i.e. it takes the value of 0). In other words, concept types are weighted, but individual query concepts can be used or discarded. The query set and relevance judgments from TREC 2014 CDS track were used to optimize concept importance weights and other parameters of the models.

2.2 Concept types

The methods that were used to obtain the 6 runs submitted to the CDS track are summarized in Table 1. Besides the type (manual or automatic) and part of the topic that they used as a query, these methods are different by the query concept types they consider.

Overall, the submitted runs utilize 4 concept sources: the query itself, PRF documents, Unified Medical Language System (UMLS) concepts extracted from the query and Google search results. Query terms, PRF documents and UMLS concepts are used by the automatic methods. For manual methods, (i.e., wsuirdma, wsuirsmb and wsuirdmb), we manually extracted a number of concepts from Google search results and added them to the expanded query, in addition to the concepts from the other 3 sources. Concept types from different sources that were used by different retrieval runs are summarized in Table 2.

All unigram concepts extracted from the original query are retained in the transformed query. Since the top retrieved documents may or may not be relevant to the original query, only a small number of unigram concepts with the

Table 1. Summary of retrieval runs submitted to TREC 2015 CDS track.

Method	Query	Method	Task
wsuirsaa	summary	automatic	A
wsuirdaa	description	automatic	A
wsuirdma	description	manual	A
wsuirsab	summary	automatic	B
wsuirsmb	summary	manual	B
wsuirdmb	description	manual	B

Table 2. Concept types utilized by submitted retrieval runs.

Concept Types	wsuirsaa	wsuirdaa	wsuirdma	wsuirsab	wsuirsmb	wsuirdmb
unigrams in topic summary	●			●	●	
ordered bigrams in UMLS concepts in topic summary	●			●	●	
unordered bigrams in UMLS concepts in topic summary	●			●	●	
unigrams in topic description		●	●			●
ordered bigrams in UMLS concepts in topic description		●	●			●
unordered bigrams in UMLS concepts in topic description		●	●			●
unigrams in PRF documents	●	●	●	●	●	●
unigrams in Google search results			●		●	●
ordered bigrams in Google search results			●		●	●
unordered bigrams in Google search results			●		●	●
unigrams in diagnosis field				●	●	●
ordered bigrams in diagnosis field				●	●	●
unordered bigrams in diagnosis field				●	●	●

highest weight in the relevance model [8] were added to the original query. The optimal number of these concepts was determined using the training data. UMLS concepts (which can consist of more than two terms) were extracted from the query using the MetaMap tool [1]. Multi-word UMLS query concepts were broken down into sequential bigrams. For example, a multi-word concept "Iron Deficiency Anemia" was represented using the Indri query language as follows:

```
1.00 #weight(
   0.40 #combine( Iron Deficiency Anemia )
   0.35 #combine( #od4( Iron Deficiency )
                  #od4( Deficiency Anemia ) )
   0.45 #combine( #uw17( Iron Deficiency )
                  #uw17( Deficiency Anemia ) )
)
```

where 0.40, 0.35 and 0.45 are the weights of the corresponding concept types. The window sizes for ordered and unordered bigrams (i.e., 4 and 17, respectively)

were determined to optimal based on the training data. It is notable that it is not necessary to normalize the mentioned weights in Indri query language to be sum-to-one as this normalization is done automatically by Indri.

Since it was shown in previous work [14] that UMLS concepts may or may not improve the performance of the medical information retrieval, only the concepts that belong to the following semantic types[1] are included in the expanded query:

- Clinical Drug
- Disease or Syndrome
- Injury or Poisoning
- Sign or Symptom
- Therapeutic or Preventive Procedure

This list was obtained from an initial list of 16 semantic types in [9] through backward elimination process [13]. Unlike [9], in which the list of considered concept types is different for each query type (i.e., "Symptom", "Diagnostic test", "Diagnosis" and "Treatment" queries), we considered the same semantic types for "Diagnosis", "Test" and "Treatment" queries.

A number of concepts that were added to the original queries in manual runs were selected from the top 10 Google search results. This selection process is done manually from the content of the documents retrieved by the Google web search engine in response to the summary or description fields of TREC CDS topics used as queries. In the case of narrative queries, the queries were modified slightly to increase the recall in Google search. Only healthcare-related concepts that are relevant to the information need of queries were added to them. The number of concepts that are extracted from Google search results and added to the transformed query depends on the relevance of documents in search results. Two factors that are considered in manually selecting the concepts from Google search results are:

1. relatedness of these concepts to the medical domain (e.g., "Kawasaki disease" is a highly related concept),
2. popularity of these concepts in medical domain (e.g., "health care" is too popular in the medical domain).

In other words, the desired concepts for query expansion in this case are the ones that are highly related to the medical domain, but not too popular.

Each concept type has different weight, as determined by its level of importance in the query. Intuitively, unigram query concepts are typically more important than unigram concepts from PRF documents. Therefore, choosing appropriate concept weights in (2) is a very important step in query transformation. We used Coordinate Ascent [12] to estimate those weights on the training data. In this optimization method, the weights are optimized one after another until convergence.

[1] `https://metamap.nlm.nih.gov/Docs/SemGroups_2013.txt`

3 Experiments

All the runs reported in this work were obtained using Indri 5.7[2] [19] IR toolkit. A two-stage document language model smoothing method proposed in [20] was used in conjunction with all retrieval models. The accuracy of 6 submitted runs in terms of Inferred Average Precision (infAP), Inferred Normalized Discounted Cumulated Gain (infNDCG), R-precision (R-prec), Precision at 10 (P@10), and Mean Average Precision (MAP) is summarized in Table 3.

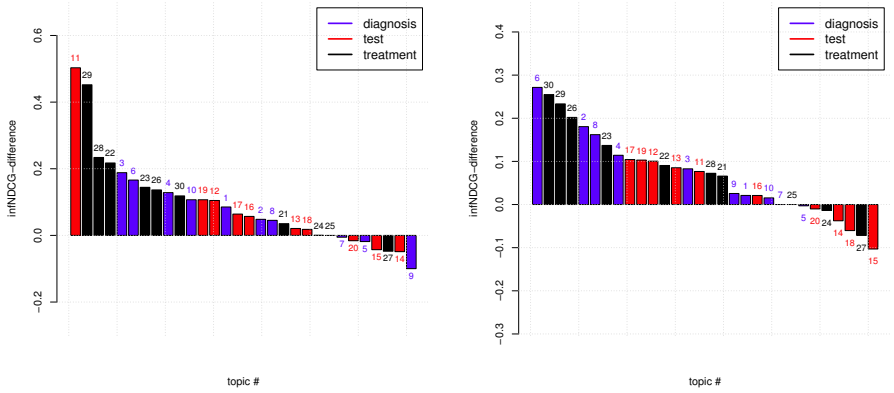
Table 3. Summary of performance for all submitted runs.

Methods	infAP	infNDCG	R-prec	P@10	MAP
wsuirsaa	0.0777	0.2928	0.2329	0.4633	0.1851
wsuirdaa	0.0842	0.2939	0.2306	0.4667	0.1864
wsuirdma	**0.0880**	**0.3109**	**0.2493**	**0.4733**	**0.1968**
wsuirsab	0.0875	0.3246	0.2656	0.5067	0.2180
wsuirsmb	0.0856	0.3208	0.2608	0.5033	0.2116
wsuirdmb	**0.1014**	**0.3690**	**0.2843**	**0.5200**	**0.2331**

Experimental results in Table 3 lead to several conclusions. First, we observe that wsuirdma, which is a manual method using unigrams from topic descriptions, PRF documents and Google search results, as well as ordered and unordered bigrams from UMLS concepts in topic descriptions and Google search results, has the highest performance in terms of all metrics for Task A of the CDS track. Second, we observe that wsuirdaa, which is an automatic method using topic descriptions as queries, outperforms wsuirsaa, which is another automatic method using topic summaries as queries. Similarly, for Task B, wsuirdmb, which is a manual method using topic descriptions as queries has significantly better retrieval accuracy than wsuirsmb, which is using topic summaries as queries. Third, we observe that incorporating information about diagnosis of the disease, which is provided in Task B, generally increases the retrieval accuracy of our models, particularly the manual ones.

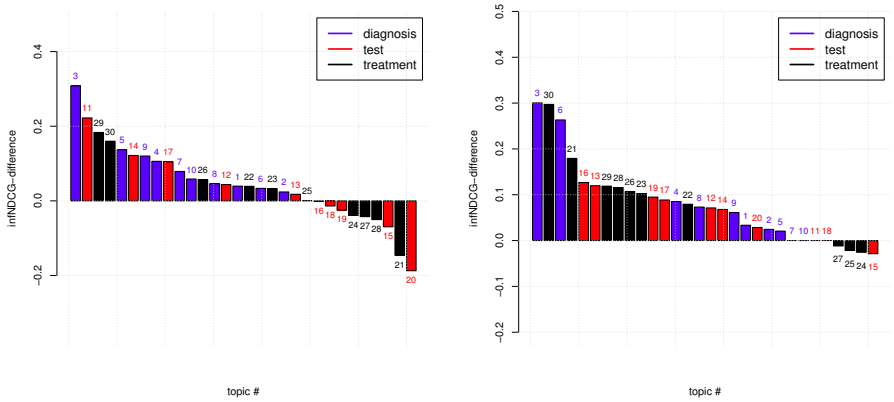
Topic-level differences in terms of infNDCG between our best automatic and manual runs and the median performance of the corresponding runs submitted to the CDS track by all other teams for Task A and Task B are illustrated in Figures 1 and 2, respectively. For Task A, our best automatic and manual runs have greater infNDCG than the median for 22 out of 30 topics (73.33%). For Task B, our best automatic run has greater infNDCG than the median for 70% of the topics, our best manual run for this task has greater infNDCG than the median for 86.67% of the topics and is slightly worse than the median for only 4 topics.

[2] `http://lemur.sourceforge.net/indri/`



(a) automatic (wsuirdaa) (b) manual (wsuirdma)

Fig. 1. Topic-level differences in terms of infNDCG between the proposed manual and automatic methods and the median for all TREC 2015 CDS track runs for Task A.



(a) automatic (wsuirsab) (b) manual (wsuirdmb)

Fig. 2. Topic-level differences in terms of infNDCG between the proposed manual and automatic methods and the median for all TREC 2015 CDS track runs for Task B.

4 Conclusion

This paper describes the participation of WSU-IR team from the Textual Data Analytics (TEANA) laboratory of Wayne State University in TREC 2015 CDS track. We provided the details of the 6 runs that our team submitted to this track, reported the retrieval performances of those runs and compared them with the median topic-level performance of all other runs submitted to this track. We observed that our best runs outperformed the median in 70% to 86% of the topics for both tasks of this track. Therefore, we conclude that adopting optimization techniques to jointly determine the weights of statistical and semantic concepts from different sources is an effective strategy for CDS retrieval models.

References

1. A. R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.
2. M. Bendersky and W. B. Croft. Discovering key concepts in verbose queries. In *Proceedings of the 31st Annual International SIGIR Conference on Research and Development in Information Retrieval*, pages 491–498, 2008.
3. M. Bendersky, D. Metzler, and W. B. Croft. Parameterized concept weighting in verbose queries. In *Proceedings of the 34th Annual International SIGIR Conference on Research and Development in Information Retrieval*, pages 605–614, 2011.
4. S. Choi, J. Choi, S. Yoo, H. Kim, and Y. Lee. Semantic concept-enriched dependence model for medical information retrieval. *Journal of Biomedical Informatics*, 47:18–27, 2014.
5. E. Erdem, E. Kocabas, H. Taylan Sekeroglu, Ö. Özgür, M. Yagmur, and T. R. Ersoz. Crystalline-like keratopathy after intravenous immunoglobulin therapy with incomplete kawasaki disease: Case report and literature review. *Case Reports in Ophthalmological Medicine*, 2013, 2013.
6. B. Koopman, P. Bruza, L. Sitbon, and M. Lawley. Towards semantic search and inference in electronic medical records: an approach using concept-based information retrieval. *The Australasian Medical Journal*, 5(9):482, 2012.
7. H. Lang, D. Metzler, B. Wang, and J.-T. Li. Improved latent concept expansion using hierarchical markov random fields. In *Proceedings of the 19th International Conference on Information and Knowledge Management*, pages 249–258, 2010.
8. V. Lavrenko and W. B. Croft. Relevance based language models. In *Proceedings of the 24th Annual International SIGIR Conference on Research and Development in Information Retrieval*, pages 120–127, 2001.
9. N. Limsopatham, C. Macdonald, and I. Ounis. Inferring conceptual relationships to improve medical records search. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, pages 1–8, 2013.
10. D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *Proceedings of the 28th Annual International SIGIR Conference on Research and Development in Information Retrieval*, pages 472–479, 2005.
11. D. Metzler and W. B. Croft. Latent concept expansion using markov random fields. In *Proceedings of the 30th Annual International SIGIR Conference on Research and Development in Information Retrieval*, pages 311–318, 2007.

12. D. Metzler and W. B. Croft. Linear feature-based models for information retrieval. *Information Retrieval*, 10(3):257–274, 2007.
13. A. Miller. *Subset selection in regression*. CRC Press, 2002.
14. W. Shen and J.-Y. Nie. Is concept mapping useful for biomedical information retrieval? In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 281–286. Springer, 2015.
15. M. S. Simpson, E. Voorhees, and W. Hersh. Overview of the trec 2014 clinical decision support track. In *Proceedings of the 23rd Text Retrieval Conference (TREC 2014)*, 2014.
16. L. Soldaini, A. Cohan, A. Yates, N. Goharian, and O. Frieder. Retrieving medical literature for clinical decision support. In *Advances in Information Retrieval*, pages 538–549. Springer, 2015.
17. P. Sondhi, J. Sun, C. Zhai, R. Sorrentino, and M. S. Kohn. Leveraging medical thesauri and physician feedback for improving medical literature retrieval for case queries. *Journal of the American Medical Informatics Association*, 19(5):851–858, 2012.
18. P. Srinivasan. Retrieval feedback in medline. *Journal of the American Medical Informatics Association*, 3:157–167, 1996.
19. T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*, volume 2, pages 2–6. Citeseer, 2005.
20. C. Zhai and J. Lafferty. Two-stage language models for information retrieval. In *Proceedings of the 25th Annual International SIGIR Conference on Research and Development in Information Retrieval*, pages 49–56, 2002.
21. N. Zhiltsov, A. Kotov, and F. Nikolaev. Fielded sequential dependence model for ad-hoc entity retrieval in the web of data. In *Proceedings of the 38th Annual International SIGIR Conference on Research and Development in Information Retrieval*, pages 253–262, 2015.
22. M. Zhong and X. Huang. Concept-based biomedical text retrieval. In *Proceedings of the 29th Annual International SIGIR Conference on Research and Development in Information Retrieval*, pages 723–724, 2006.