TUW @ TREC Clinical Decision Support Track 2015

João Palotti and Allan Hanbury

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

Abstract. In this document, we describe the participation of Vienna University of Technology (TUW in German) in both tasks A and B of the TREC Clinical Decision Support (TREC-CDS) Track 2015. Based on the 2014 data, we concluded that query expansion with PRF resulted in large improvements over a BM25 baseline. Thus, we investigate a manner to add an intermediary layer based on a subset of the concepts annotated by MetaMap. This acts as a way to add weight to relevant concepts in the query and slightly expand the query with the preferred name of relevant concepts, before performing the query expansion with PRF. For TREC-CDS 2014, our method could reach a precision at 10 (P@10) of 0.40, while the best result of that year was 0.39. For 2015, we could reach a P@10 of 0.41 using the intermediary layer proposed, a small improvement over our baseline of P@10 of 0.39 when using only the original query expanded with PRF.

Keywords: Medical Information Retrieval, Evaluation

1 Introduction

It is estimated that more than 80% of the American population uses the Web to seek health information [1]. Small wonder that it attracts attention of the Information Retrieval community, as a significant improvement could mean millions of more satisfied users. To develop tools for health search and foster research in this area, a number of shared tasks are yearly organized. Among those, there was the TREC Genomics Track [14] which ran from 2003 to 2007, the TREC Medical Records Track [16] running in 2011 and 2012, the ImageCLEFmed Track on medical image retrieval [5,6] running between 2003 and 2013, and the ShARe/CLEF eHealth Evaluation Lab [15,3,11,12] running since 2013. Here we briefly describe the participation of Vienna University of Technology on the second TREC Clinical Decision Support Track (TREC-CDS).

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. A complete description of the task can be found in [13] or online at trec-cds.org/2015.html.

Our Contribution

A traditional approach in IR is the use of query expansion with pseudo-relevance feedback (PRF). It is usually done in a two-step procedure: first a query is issued to the search engine which returns the most relevant/similar documents in the collection, then important/discriminative terms are drawn from the top documents, and added back to the original query, which is issued again. The idea is that relevant terms could be added to the original query, potentially refining the final results.

We propose to add a domain dependent query expansion layer before the PRF step. We map the user query to concepts in large domain specific vocabulary, such as Unified Medical Language System (UMLS), and extract the preferred names for the concepts. We selected some concepts based on their semantic type, keeping only the ones that belong to a shortlist of relevant types, representing diseases, symptoms, remedies or group ages. We empirically attributed different weights for the part of the query which triggered a concept, as well as to the preferred name of the concept, potentially making the query biased to some concepts. A detailed description of our approach with a running example is given in the following section.

2 Approach

Given a topic, as shown in Figure 1, MetaMap was used to map the summary field of each topic to UMLS concepts, as show in Table 1. We employed MetaMap with its default settings, and we allowed it to map the same text to multiple concepts. For example 'year' is mapped to two concepts, both with the semantic type 'tmco', which stands for *'Temporal Concept'*. As show in Table 1, usually MetaMap has a high recall, but a low precision. To cope with that, and have a smaller selection of highly relevant terms for query expansion, we selected a number of semantic types that we were relevant for the task, ignoring all mappings to other semantic types. The shortlisted semantic types used are listed in the Table 2. They were inspired by past work, in which the semantic types were employed to define classes such as "Symptom", "Remedy", or "Disease" [7,2,8,10,9]. Some additional types were used as well, because they could be relevant for the task, such as 'aggp' (Age Group).

As shown in Table 1, we are specially interested in knowing which part from the original topic was used as a trigger for a mapping, as well as the preferred name of mapped concept. We do not use the concept ID itself, but the preferred name of the concept instead. Our goal is attributing more weight to the important concepts detected in the original query (trigger text), as well as to the preferred name for that concept, which sometimes might differ from the trigger text, allowing us to explore related terms as well. We define $OWeight$ as the weight for the original terms in the query, $TWeight$ as the weight attributed to the triggers of a mapping, $PWeight$ as the weight attributed to the preferred name of a concept mapped from the text, and $DWeight$ as the weight given to the diagnose (disease name) in Task-B scenario (see the task description at

trec-cds.org/2015.html). We explore different values for the weights based on the results from data of TREC-CDS 2014, and the values chosen for each of our runs are shown in the Table 3.

Finally, after expanding the queries with terms from MetaMap, we use Bose-Einstein 1 (BO1) to perform a pseudo-relevance feedback, adding even more terms to the original query. Figure 2 shows all the steps done for topic 1 of TREC-CDS 2014, with $OWeight = 3.0$, $TWeight = 2.0$, and $PWeight = 1.0$.

Specially for TaskB, we took advantage of the diagnose provided by the organizers, annotating it with MetaMap, and attributing more weight for the concepts generated by this important mapping.

```
<topic number="1" type="diagnosis">
  <description>....</description>
  <summary>58-year-old woman with hypertension and obesity presents with
    exercise-related episodic chest pain radiating to the back.
  </summary>
</topic>
```

Fig. 1: Topic 1 from TREC-CDS 2014

Table 1: Some of the mappings trigged by the Topic 1 of TREC-CDS 2014.

Trigger	**Concept**	**Preferred Name**	**Symtype**
year	C0439508	per year	tmco
year	C0439234	year	tmco
old	C0580836	old	tmco
hypertension	C1963138	hypertension adverse event	fndg
hypertension	C0020538	hypertensive disease	dsyn
obesity	C1963185	obesity adverse event	fndg
...	...	...	...
exercise	C0015259	exercise	dora
...	...	...	...
back	C1995000	back structure, excluding neck	blor
back	C0460009	back structure, including back of neck	blor

3 Results

Tables 4 and 5 show the results respectively for Precision at 10 (P@10) and R-Precision (R-prec) obtained by our approach for 2014 and 2015 tasks. All the experiments were performed using Terrier 4.0 as search system. BM25 is a

Table 2: Semantic types used in the MetaMap filtering step. Check https://metamap.nlm.nih.gov/Docs/SemanticTypes_2013AA.txt for an explanation of the meaning of each type

aggp	chvf	eico	modb	phsu
anab	chvs	emod	neop	rcpt
antb	clna	enzy	nnon	sosy
bacs	clnd	fndg	nsba	strd
bodm	comd	horm	opco	topp
bpoc	diap	irda	orch	vita
carb	dsyn	lipd	patf	

Table 3: Different weights were used for each run submitted by TUW. All runs are based on the BM25 retrieval model and use BO1 for PRF. Runs TUW1 and TUW4 do not use the MetaMap mappings. Runs TUW5 and TUW6 used two different TWeight and PWeight: the first one is the same as TUW2 and TUW3, the second one is a special weight for the mappings provided by the diagnose term added in Task-B.

Task	RunID	OWeigth	TWeight	PWeight	DWeight
A	**TUW1**	1	-	-	-
	TUW2	1	1	1	-
	TUW3	3	2	1	-
B	**TUW4**	1	-	-	6
	TUW5	1	1, 4	1, 4	6
	TUW6	3	2, 5	1, 4	6

```
Original Text:
==============
58-year-old woman with hypertension and obesity presents with exercise-related
episodic chest pain radiating to the back.

After expanding with MetaMap (OWeight = 3.0, TWeight = 2.0, PWeight = 1.0)
==========================================================================
58^3.0 year^3.0 old^3.0 woman^3.0 with^3.0 hypertension^3.0 hypertension^1.0
hypertension^2.0 disease^2.0 hypertensive^2.0 radiating^1.000000 the^3.000000
and^3.0 obesity^3.0 obesity^2.0 obesity^1.0 adverse^2.0 event^2.0 presents^3.0
exercise^3.0 exercise^2.0 exercise^1.0 management^2.0 related^3.0 related^2.0
personal^2.0 relate^2.0 related^1.0 resin^2.0 vinyl^2.0 status^2.0 episodic^3.0
chest^3.0 chest^2.0 chest^1.0 pain^3.0 pain^2.0 pain^1.0 radiating^3.0
radiating^2.0 to^3.0 back^3.0

Final Query after PRF with BO1 (3 documents and 10 terms)
=========================================================
hypertens^1.056125896 radiat^0.750000000 exercis^0.835373747 pain^0.862381156
old^0.375000000 resin^0.250000000 chest^0.928651999 year^0.375000000 58^0.375000000
person^0.250000000 obes^0.750000000 event^0.250000000 diseas^0.250000000
episod^0.426995833 manag^0.250000000 statu^0.250000000 present^0.375000000
back^0.375000000 advers^0.250000000 woman^0.375000000 vinyl^0.250000000
angina^0.396873208 aortic^0.284881163 dissect^0.233308385 coronari^0.164579531
ischemia^0.154572160 myocardi^0.146660196 ecg^0.096813983 cardiomyopathi^0.092673589
diaphoresi^0.075201212 arteri^0.074352515 heart^0.066933405 st^0.064917927
obstruct^0.059847174 cardiac^0.055418674 substern^0.053208445 infarct^0.051212615
```

Fig. 2: Modifications made for topic 1 of TREC-CDS 2014: from the original summary of topic 1 to the actually issued query

weak baseline not submitted, used here only for comparison. TUW1 and TUW4 are runs using BM25 and pseudo-relevance feedback, TUW2 and TUW5 are runs using MetaMap for query expansion without assigning any weights for the terms followed by PRF with BO1 (adding 10 terms from the top 3 documents). Finally, different weights are assigned for the terms annotated by MetaMap for runs TUW3 and TUW6. PRF with BO1 (3 documents, 10 terms) is also performed. As it is shown in both tables, the approach of TUW3 and TUW6 obtained very high gains in terms of P@10, when it is compared to the pure BM25 (33% improvement) or even BM25 with PRF (13% improvement). Same trend for R-Precision, with 46% of improvement over the pure BM25 and 9% of improvement over the BM25 with PRF.

For 2015, a larger gain was expected. However, we just had a gain of 5% from TUW3 compared to TUW1 for P@10 on Task A, and a loss of 3% from TUW2 to TUW1 for P@10 on the same task. For Task B, the trend was different, with TUW5 having the best results, 3% better than TUW4. In Table 6, all official measures are shown, including infnDCG and infAP. Note that for inferred measures, there was no gain in adding a step of query expansion with MetaMap. Additionally, Precision at 10 and R-Precision results for each topic are shown in Figures 3 and 4.

Table 4: Results for Precision at 10

RunIDs	**Variation**	**2014**	**2015-A**	**2015-B**
-	Higher Score	0.39 [4]	-	-
-	Median	0.23	0.34	0.45
-	BM25	0.30	0.36	0.46
TUW 1&4	BM25 + PRF	0.34	0.39	0.52
TUW 2&5	BM25 + MetaMap + PRF	0.36	0.38	0.54
TUW 3&6	BM25 + WMetaMap + PRF	0.40	0.41	0.51

Table 5: Results for R-prec

RunIDs	**Variation**	**2014**	**2015-A**	**2015-B**
-	Higher Score	0.22 [17]	-	-
-	Median	0.13	0.16	0.21
-	BM25	0.15	0.17	0.24
TUW 1&4	BM25 + PRF	0.20	0.20	0.26
TUW 2&5	BM25 + MetaMap + PRF	0.19	0.19	0.27
TUW 3&6	BM25 + WMetaMap + PRF	0.22	0.19	0.26

Table 6: Official results for TREC-CDS 2015

Task	Runs	P@10	InfnDCG	infAP	R-Prec
A	**Avg. Best**	0.68	0.44	0.13	0.32
	Avg. Median	0.34	0.20	0.04	0.16
	TUW1	0.39	0.24	0.06	0.20
	TUW2	0.38	0.22	0.05	0.19
	TUW3	0.41	0.23	0.05	0.19
B	**Avg. Best**	0.78	0.53	0.17	0.39
	Avg. Median	0.45	0.28	0.06	0.21
	TUW4	0.52	0.36	0.10	0.26
	TUW5	0.54	0.36	0.10	0.27
	TUW6	0.51	0.34	0.09	0.26

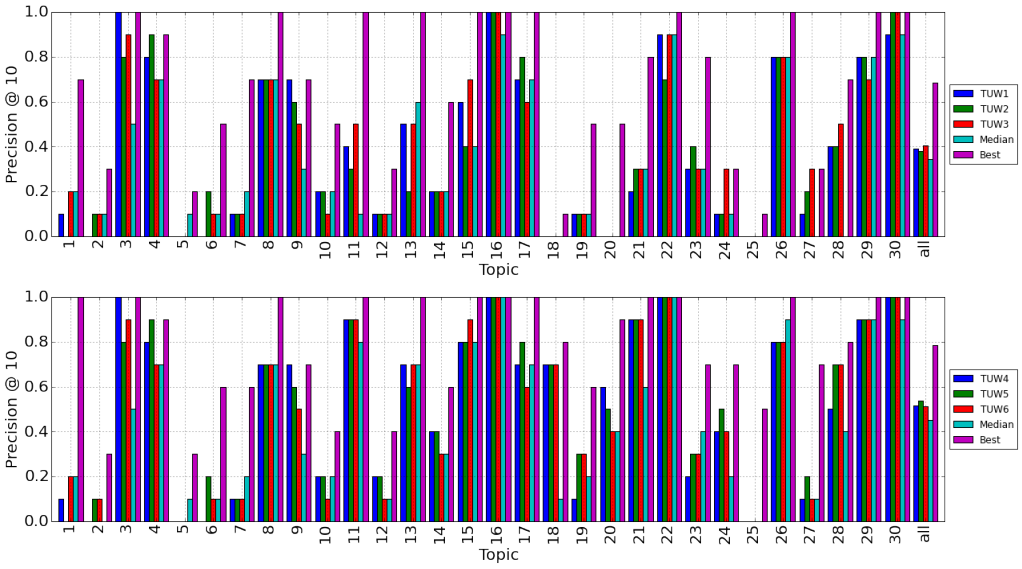

Fig. 3: Precision at 10 for each topic for TREC-CDS 2015 tasks A and B

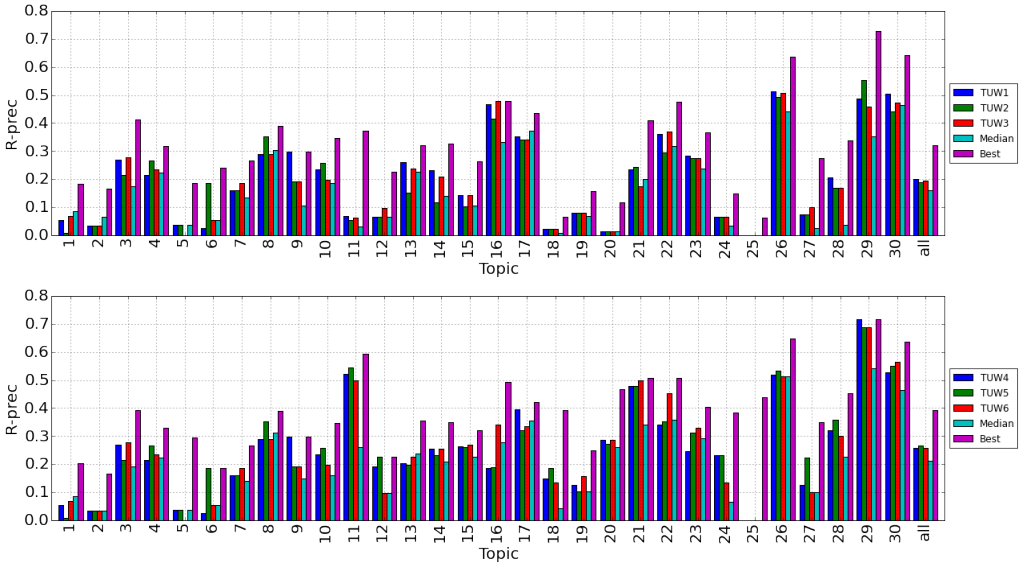

Fig. 4: R-Precision for each topic for TREC-CDS 2015 tasks A and B

4 Conclusion

In this paper we report on the experiments made by Vienna University of Technology (TUW) for TREC-CDS. We created a simple, but effective method based on query expansion of mapped terms from the original query, before applying PRF. Our results have shown large improvements for TREC-CDS 2014, however it did not result in large improvements for TREC-CDS 2015 as well. A further detailed investigation need to be conducted to understand why the intermediary layer worked so well for 2014 and not so well for 2015.

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. S. Fox. *Health topics: 80% of internet users look for health information online.* Pew Internet & American Life Project, 2011.
2. A. S. Jadhav, A. P. Sheth, and J. Pathak. Online information searching for cardiovascular diseases: An analysis of mayo clinic search query logs. *Studies in Health Technology and Informatics*, pages 702–706, 2014.

3. L. Kelly, L. Goeuriot, H. Suominen, T. Schreck, G. Leroy, D. L. Mowery, S. Velupillai, W. W. Chapman, D. Martínez, G. Zuccon, and J. 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. A. Mourão, F. Martins, and J. Magalhães. Novasearch at TREC 2014 clinical decision support track. In *Proceedings of TREC*, 2014.
5. H. Müller, P. Clough, T. Deselaers, and B. Caputo. *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*. Springer Publishing Company, Incorporated, 1st edition, 2010.
6. H. Müller, A. García Seco de Herrera, J. Kalpathy-Cramer, D. Demner-Fushman, S. Antani, and I. Eggel. Overview of the imageclef 2012 medical image retrieval and classification tasks. In *CLEF 2012 working notes*, 2012.
7. A. Névéol, R. I. Dogan, and Z. Lu. Semi-automatic semantic annotation of pubmed queries: A study on quality, efficiency, satisfaction. *Journal of Biomedical Informatics*, 44(2):310–318, 2011.
8. J. Palotti, A. Hanbury, and H. Muller. Exploiting health related features to infer user expertise in the medical domain. In *Proceedings of WSCD Workshop on Web Search and Data Mining*. John Wiley & Sons, Inc., 2014.
9. J. Palotti, A. Hanbury, H. Muller, and C. K. Jr. How users search and what they search for in the medical domain. *Information Retrieval Journal*, 2015.
10. J. Palotti, V. Stefanov, and A. Hanbury. User intent behind medical queries: An evaluation of entity mapping approaches with metamap and freebase. In *Proceedings of the 5th Information Interaction in Context Symposium*, IIiX '14, pages 283–286. ACM, 2014.
11. J. Palotti, G. Zuccon, L. Goeuriot, L. Kelly, A. Hanbury, G. Jones, M. Lupu, and P. Pecina. Clef ehealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. In *CLEF 2015 Online Working Notes*. CEUR-WS, 2015.
12. J. Palotti, G. Zuccon, L. Goeuriot, L. Kelly, A. Hanburyn, G. J. Jones, M. Lupu, and P. Pecina. CLEF eHealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. In *CLEF 2015 Online Working Notes*. CEUR-WS, 2015.
13. K. Roberts, M. Simpson, D. Demner-Fushman, E. Voorhees, and W. Hersh. State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the trec 2014 cds track. *Information Retrieval*, 7 2015.
14. P. M. Roberts, A. M. Cohen, and W. 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.
15. H. Suominen, S. Salanterä, S. Velupillai, W. W. Chapman, G. K. Savova, N. Elhadad, S. Pradhan, B. R. South, D. L. Mowery, G. J. F. Jones, J. Leveling, L. Kelly, L. Goeuriot, D. Martínez, and G. 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.
16. E. 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.
17. T. Xu, D. W. Oard, and P. McNamee. HLTCOE at TREC 2014: Microblog and clinical decision support. In *Proceedings of TREC*, 2014.