DCU@TRECMed 2012: Using ad-hoc baselines for domain-specific retrieval

Johannes Leveling, Lorraine Goeuriot, Liadh Kelly, Gareth J. F. Jones

Centre for Next Generation Localisation (CNGL), School of Computing,
Dublin City University, Dublin 9, Ireland
firstname.lastname@computing.dcu.ie

Abstract. This paper describes the first participation of DCU in the TREC Medical Records Track (TRECMed). We performed some initial experiments on the the 2011 TRECMed data based on the BM25 retrieval model. Surprisingly, we found that the standard BM25 model with default parameters, performs comparable to the best automatic runs submitted to TRECMed 2011 and would have resulted in rank four out of 29 participating groups. We expected that some form of domain adaptation would increase performance. However, results on the 2011 data proved otherwise: concept-based query expansion decreased performance, and filtering and reranking by term proximity also decreased performance slightly. We submitted four runs based on the BM25 retrieval model to TRECMed 2012 using standard BM25, standard query expansion, result filtering, and concept-based query expansion. Official results for 2012 confirm that domain-specific knowledge does not increase performance compared to the BM25 baseline as applied by us.

1 Introduction

This paper describes the first participation of DCU in the TREC Medical Records Track (TRECMed). TRECMed is an instance of domain-specific information retrieval (IR) and ran for the first time in 2011, with 29 participating groups. A review of the 2011 participants' approaches shows that the most successful approaches include:

- Information extraction on the corpus: applying natural language processing (NLP) techniques (e.g. chunking, PoS tagging, lemmatization) [1–3]; linking with additional concept bases, or medical ontologies [4, 1, 5, 2, 3]; expanding ICD9 codes (International Statistical Classification of Diseases and Related Health Problems, version 9) for the patient's admission or discharge status [1, 5]; treating negation (e.g. negative test results or symptoms) [4, 1, 5, 2, 3];
- Query expansion based on external knowledge (e.g. medical web sites, knowledge bases or Wikipedia) [2, 3];
- Result filtering based on extracted patient features such as ethnicity, age, and gender [4, 1, 5];
- A popular IR framework used in the 2011 campaign is the Lucene toolkit (and its default retrieval model) [4, 1, 5, 2, 3].

Our goal for the participation in TRECMed was to investigate query processing and different expansion techniques while trying to establish a good baseline for this task.

2 Related Work

We view medical record retrieval as an instance of domain-specific IR. There have been several evaluation tasks in IR evaluation campaigns such as TREC, NTCIR, and CLEF which focus on domain-specific IR, e.g. TREC-Chem[1] [6], patent retrieval[2] [7], or geographic IR[3] [8].

Domain adaptation for IR has not proven to be consistently successful. In GIRT, the domain-specific IR task at CLEF, few participants used meta-data such as additional document fields containing subject terms or a domain-specific thesaurus [9, 10] as standard IR models yielded a high performance. Similarly, one important result from evaluation of geographic IR (GIR) is that adding large gazetteers with geographic knowledge decreases performance. For GIR, simple text-based retrieval (with a bag-of-words approach) turned out to be a very strong baseline [11].

Armstrong et al. [12] analysed several years of experiments on TREC data and found that very few results reported in the literature outperform strong baseline experiments. Most experiments claim a significant improvement, but over a weak baseline and not over the best results on the same data. Thus, improvements on the data do not add up, as it becomes more difficult to improve on good results.

3 System Description

The objective of our participation in TRECMed 2012 was to establish a baseline BM25 system and compare different query expansion techniques. Our system employs approaches that were described as successful by last year's participants and comprises simple preprocessing and analysis steps:

- The document terms in an initial indexing run were manually examined and a list of misspelled and run-together words was compiled. This list was used to correct terms in the final indexing stage.
- A single text index for fields was employed, formed from the report text and the textual description of the ICD9 fields.
- All report documents are indexed separately, i.e. retrieved results have to be mapped to patient visits. The system then returns the document with the maximum score to map reports to visits. This approach was found to best the best way to map retrieved documents to visits [3].

[1] http://wiki.ir-facility.org/index.php/TREC_Chemistry_Track
[2] http://www.cl.cs.titech.ac.jp/~fujii/ntc8patmt/rs.html
[3] http://www.linguateca.pt/GikiCLEF/

- Retrieval and query expansion are based on the BM25 model [13].
- An additional filtering step filters results by applying constraints from the query pertaining to the patient's age, ethnicity, gender, or admission status to the result set. Similar approaches have been investigated for TRECMed 2011 [1, 5, 2].

3.1 Document preprocessing

All report files (which pertain to a patient's visit) were indexed as separate documents. ICD9 codes were mapped to a description of the code, usually a short phrase/sentence. For instance, code *253.5* corresponds to the disease *Diabetes insipidus*. They were then stored in the additional document fields ICD9_DIS_DIAGNOSIS_TEXT and ICD9_ADM_DIAGNOSIS_TEXT. The fields REPORT_TEXT, TYPE, SUBTYPE, ICD9_DIS_DIAGNOSIS_TEXT, and ICD9_ADM_DIAGNOSIS_TEXT were used to create a single index for the body of text.

3.2 Spelling correction

Spelling errors may have a detrimental impact on the system's performance. We manually corrected index terms by examining all index terms from an initial run. Corrections were only added to the list when the correction was unambiguous. The correction was not performed by a medical expert, so many incorrect technical terms may have been missed. Even so, we compiled a list of 9533 spelling errors from the medical documents, which was added to a list of 4192 frequent spelling errors compiled from Wikipedia. During indexing, misspelled words are replaced with their corrections from this list. As an example, we found eight misspellings for the word *admission*: *addmision*, *admision*, *admissin*, *admissoin* *admisson*, *admisssion*, *admsission*, *dmission*.

3.3 Retrieval

Most implementations of BM25 for Lucene approximate document length as the number of characters, approximate field length by the maximum field length (for BM25F), or store the length information with a loss of precision (e.g. [14]). This can result in lower performance and/or different optimal model parameters. We employ our own BM25 implementation for Lucene which follows the original BM25 description [13] closely. Our system employed Lucene's standard tokenization and a standard stopword list containing 33 stopwords.

3.4 Query expansion

We applied two approaches to query expansion: the standard approach described by Robertson et al. [13] and concept expansion.

- Query expansion: In the default query expansion, terms from the top ranked documents are ranked by a term selection value [15] and the top R terms are added to the query. For our experiments, 10 terms were extracted from the top 10 documents.
- Concept expansion: We annotated the queries with concepts from the UMLS (Unified Medical Language System) thesaurus[4], using the Metamap system [16]. For each phrase, the system gives a ranked list of potential mapping concepts (called *Meta Candidates*) and one or several (in case of equal scores) mapped concepts (called *Meta Mapping*). We used the *Meta Mapping* concept list and their short description to extend the query, for example: *Patients with complicated GERD who receive endoscopy* will be extended with *Gastroesophageal reflux disease, Clinic / Center - Endoscopy*

3.5 Result filtering

Initial retrieval results were filtered with respect to constraints given in the query regarding the age, gender, ethnicity, and admission status of a patient. Sentences containing the anonymized age information of a patient (***AGE**) were extracted from the document collection and manually annotated to obtain annotation patterns. Roughly 500 patterns were extracted. For example, the word sequence *"is an ** age [in 80s] yr old wm admitted"* allows to infer that an 80-89 year old white male patient was admitted. The longest match between sentences in a document and the patterns was then used to augment the document's meta-data and overwrite the default value (e.g. *unknown*) for the *age*, *gender*, *ethnicity*, and *admission status* features. For the filtering step, only documents with exactly matching values were kept, while allowing the value *unknown* to match with any value. Table 1(a), 1(b), 1(c), and 1(d) show the distribution of values in the annotated document collection.

4 Experiments and Analysis

4.1 Experiments on 2011 Data

We performed initial experiments on the the 2011 TRECMed data based on the previously described setup. We tested four runs on the 2011 data, using

i) standard BM25 retrieval,
ii) i) + standard query expansion (QE),
iii) ii) + result filtering, and
iv) ii) + concept-based query expansion (CE).

Results are shown in Table 2. Surprisingly, we found that our baseline experiment, applying the standard BM25 model with default parameters, performs comparable to the best automatic runs submitted to TRECMed 2011. It would

[4] http://www.nlm.nih.gov/research/umls/

Table 1. Distribution for extracted patient features.

(a) *age*

age	count
0-12	37
13-20	529
20-29	2.684
30-39	2.675
40-49	5.385
50-59	6.611
60-69	6.561
70-79	6.661
80-89	6.418
90+	788
unknown	57.352
Σ	95.701

(b) *gender*

gender	count
female	16.824
male	14.592
unknown	64.285
Σ	95.701

(c) *ethnicity*

ethnicity	count
asian	18
black	1.400
hispanic	4
white	5.885
unknown	88.394
Σ	95.701

(d) *admission status*

admission status	count
admitted	12.369
not admitted	8.222
unknown	75.110
Σ	95.701

have ranked among the top five out of 29 participating groups (0.4052 MAP, 0.5082 bpref, 0.6 P@10). Evaluation results for the best automatic runs range from 0.552-0.494 bpref, 0.656-0.568 P@10, and 0.440-0.401 Rprec for the top five participants in TRECMed 2011 [17].

Domain-specific IR typically requires adaptation of at least one component of a search system to the domain, e.g. by including domain-knowledge from ontologies or modifying the retrieval model. We expected that some form of domain adaptation would increase performance compared to the BM25 retrieval baseline. However, results on the 2011 data did not confirm this: standard query expansion decreases MAP and bpref, but slightly increased precision at early ranks; concept-based query expansion decreased performance in general, and filtering and reranking results also decreased performance.

Table 2. Results on 2011 topics.

Run	Description	MAP	bpref	P@10
i)	BM25	0.4052	0.5082	0.6000
ii)	BM25+QE	0.3249	0.4867	0.4853
iii)	BM25+QE+filter	0.3229	0.4857	0.4824
iv)	BM25+CE+filter	0.3425	0.5116	0.4882

4.2 Experiments on 2012 Data

Results for our four official submitted runs are shown in Table 3. We expected that, as for the 2011 data, concept-based query expansion and result filtering will decrease performance (bpref and MAP) significantly, compared to the simple BM25 baseline. We observe that performance in general is much lower compared to results on 2011 data, which may be due to more difficult topics. However, comparing our own experiments, the expected decreases is not as high as on 2011 data, e.g. run iii) vs. run i).

Table 3. Results on 2012 topics.

Run	Description	MAP	bpref	P@10
i)	BM25	0.2930	0.3462	0.4638
ii)	BM25+QE	0.2562	0.3163	0.4213
iii)	BM25+QE+filter	0.2734	0.3331	0.4553
iv)	BM25+CE+filter	0.2552	0.3152	0.4191

We see several possible explanations: 1. Our approach to include medical knowledge performs worse than the approaches of other participants because our annotation is less accurate. Using additional domain information with low accuracy degrades the performance. However, this would not explain why BM25 with default settings performs exceptionally well. 2. The BM25 retrieval model is superior to Lucene's internal ranking scheme, which is a variant of tf-idf with support for boosting terms and documents. BM25 can still be considered a strong baseline, even for domain-specific IR and twenty years after its introduction. Lucene (with its standard ranking model) was used by many of the top performing groups in TRECMed 2011.

5 Conclusion

Including domain-specific adaptation results in more complex indexing and retrieval workflows, but intuitively, adaptation should result in a significant performance improvement over the standard retrieval baseline. For ad-hoc IR, Armstrong et al. [12] pointed out that comparing against a weak baseline allows observation of a significant performance increase that cannot be replicated against a strong baseline. We argue that for a domain-specific task, strong baselines are needed even more to isolate domain-adaptation issues and make their effects observable. Strong generic baselines can be derived from open-domain retrieval (i.e. ad-hoc IR). Weak baselines are not adequate and invalidate conclusions on the effect of domain adaptation, because an improvement over a weak baseline is harder to reproduce over a stronger baseline.

We would like to propose that additional meta-data or annotations are made available by participants in evaluation tasks as stand-off annotations for the document collection so that participating groups can perform experiments on the same meta-data (e.g. document ID and extracted patient features). This would lower the entry-level for new participants and make results between participants more comparable, as the quality of generating additional meta-data would be separate from the quality of using it.

Acknowledgment

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project at DCU and by funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n°257528 (KHRESMOI).

References

1. King, B., Wang, L., Provalov, I., Zhou, J.: Cengage Learning at TREC 2011 medical track. In: TREC 2010, Gaithersburg, Maryland, National Institute of Standards and Technology (NIST) (2011)
2. Goodwin, T., Rink, B., Roberts, K., Harabagiu, S.M.: Cohort shepherd: discovering cohort traits from hospital visits. In: TREC 2010, Gaithersburg, Maryland, National Institute of Standards and Technology (NIST) (2011)
3. Schuemie, M., Trieschnigg, D., Meij, E.: DutchHatTrick: Semantic query modeling, ConText, session detection, and match score maximization. In: TREC 2010, Gaithersburg, Maryland, National Institute of Standards and Technology (NIST) (2011)
4. Demner-Fushman, D., Abhyankar, S., Jimeno-Yepes, A., Loane, R., Rance, B., Lang, F., Ide, N., Apostolova, E., Aronson, A.R.: A knowledge-based approach to medical records retrieval. In: TREC 2010, Gaithersburg, Maryland, National Institute of Standards and Technology (NIST) (2011)
5. Gurulingappa, H., Müller, B., Hofmann-Apitius, M., Fluck, J.: A semantic platform for information retrieval from e-health records. In: TREC 2010, Gaithersburg, Maryland, National Institute of Standards and Technology (NIST) (2011)
6. Lupu, M., Piroi, F., Huang, X., Zhu, J., Tait, J.: Overview of the TREC 2009 Chemical IR Track. In Voorhees, E.M., Buckland, L.P., eds.: Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009. Volume Special Publication 500-278., NIST (2009)
7. Fujii, A., Iwayama, M., Kando, N.: Overview of the patent retrieval task at the NTCIR-6 workshop. In: Proceedings of NTCIR-6 Workshop Meeting. (2007) 359–365
8. Santos, D., Cardoso, N., Carvalho, P., Dornescu, I., Hartrumpf, S., Leveling, J., Skalban, Y.: GikiP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia. In: Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers. Volume 5706 of LNCS., Springer (2009) 894–905
9. Kluck, M., Stempfhuber, M.: Domain-specific track CLEF 2005: Overview of results and approaches, remarks on the assessment analysis. In: Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evalution Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers. Volume 4022 of LNCS., Springer (2006) 212–221

10. Petras, V., Baerisch, S., Stempfhuber, M.: The domain-specific track at CLEF 2007. In: Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers. Volume 5152 of LNCS., Springer (2008) 160–173
11. Mandl, T., Carvalho, P., Nunzio, G.M.D., Gey, F.C., Larson, R.R., Santos, D., Womser-Hacker, C.: GeoCLEF 2008: The CLEF 2008 cross-language geographic information retrieval track overview. In: Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers. Volume 5706 of Lecture Notes in Computer Science., Springer (2009) 808–821
12. Armstrong, T.G., Moffat, A., Webber, W., Zobel, J.: Improvements that don't add up: ad-hoc retrieval results since 1998. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, ACM (2009) 601–610
13. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M.: Okapi at TREC-3. In Harman, D.K., ed.: Overview of the Third Text Retrieval Conference (TREC-3), Gaithersburg, MD, USA, National Institute of Standards and Technology (NIST) (1995) 109–126
14. Pérez-Iglesias, J., Pérez-Agüera, J.R., Fresno, V., Feinstein, Y.Z.: Integrating the Probabilistic Models BM25/BM25F into Lucene. CoRR **abs/0911.5046** (2009)
15. Walker, S., Robertson, S.E., Boughanem, M., Jones, G.J.F., Jones, K.S.: Okapi at TREC-6 Automatic ad hoc, VLC, routing, filtering and QSDR. In: TREC-6. (1997) 125–136
16. Aronson, A.R., Lang, F.M.: An overview of metamap: historical perspective and recent advances. Journal of the American Medical Informatics Association **17** (2010) 229–236
17. Voorhees, E.M., Tong, R.M.: Overview of the TREC 2011 medical records track. In: TREC 2010, Gaithersburg, Maryland, National Institute of Standards and Technology (NIST) (2011)