# Search for Clinical Records: RMIT at TREC 2011 Medical Track

Iman Amini* Mark Sanderson* David Martinez† Xiaodong Li*
*RMIT Dept of Computer Science and NICTA, Australia
†NICTA and the University of Melbourne, Dept of CSSE, Australia
iman.amini,mark.sanderson,xiaodong.li@rmit.edu.au
david.martinez@nicta.edu.au

## Abstract

We combine several techniques to participate in Medical TREC 2011 and later we decompose our combined methodologies to gain a thorough understanding of the effects of each individual technique. In this paper we focus on Information Extraction and Expansion to find the best setting for an ideal IR system. Results suggest that Information Expansion is a key strategy in finding relevant reports for a medical query.

## 1 Introduction

TREC has been a key venue in testing and evaluating IR systems against comprehensive and realistic test collections. Medical TREC has kicked off for the first time this year [1]. Prior to Medical TREC, IR community had a minimal access to clinical data, however, medical context had more popularity amongst the Natural Language Processing (NLP) researchers. Plenty of IR techniques that have been tested successfully in general IR on large collections such as clueweb and legal track, are yet to be evaluated on the medical dataset [4].

RMIT submitted four distinct runs to Medical TREC three of which were used for pooling and therefore judged. In this paper we outline the methods and experiments performed and report the final results. Prior to the release of final results we performed our own manual judgments and produced a preliminary evaluation table outlined in this paper. We conclude the paper by showing a comparison table of RMIT's best run with the arithmetic mean of median and the best submissions.

## 2 Analysis of the Collection

Approximately 100,000 clinical reports an aggregation of overall 17,000 individual visits were gathered during the year 2007 from the U.S hospitals and used for Medical track. The collection is fairly small comparing to Clueweb and other well known IR collections but yet it is the biggest existing medical/clinical test collection.

We started off by parsing and analysing the collection to gain familiarity with the structure of the collection and to identify salient as well as non informative chunks of the text. We found the reports to be fairly dense in terms of the vocabulary used in comparison to the rich vocabulary used in medical/clinical papers. Also in contrast to our pre-assumption reports do not show spelling mistakes, saving a huge amount of data cleansing.

[1] TREC Conference Home Page:http://trec.nist.gov/

However, to this end, we are not sure about the existence of semantic mistakes such as assigning a wrong medical code to a patient's report. International Codes/Concepts for Diseases (ICD) could be located in two sections of the text: discharge and admit diagnosis, however, these codes needed to be mapped into textual descriptions in order to match the language of the topics.

Using simple pattern matching we extracted section headings and identified segments pertaining to different population and age groups. We found that 12,006 reports had one visit associated while 2,387 of the reports had more than or equal to 10 visits.

# 3 Relevance Judgment Early on

We manually judged the relevancy of retrieved documents using depth-k for k=20 for the retrieved documents and also simulated the automatic evaluation benchmark proposed in [9] for the queries that contained diagnostic codes.

Due to the lack of evidence judgment for the given 4 sample queries we gathered a pool of documents retrieved by three different systems, these systems differed only in their ranking algorithms.

Annotation was done by two authors of this paper and it turned out that we had extremely low agreement on two topics, after receiving guidelines from a medical student and looking through the manual annotation, however, we arrived at perfect agreement.

Two sample queries over which we had disagreement are given below:

- Patients treated for lower extremity chronic wound
- Elderly patients with ventilator-associated pneumonia

After speaking to a medical student we found out that one of the annotators mistakenly marked all types of wounds as relevant and for the second query the term associated is apparently important when used in the phrase and not individually.

# 4 Query Expansion

Our aim was to automatically identify terms that needed expansion and choose the best external knowledge source to expand the queries with.

We find that often query terms added from Wikipedia lead to over expansion and later we address this issue by using DBpedia[2]. which offers a more structured way to access the content of Wikipedia, however, we eliminate non informative terms and headings extracted from the DBpedia.

Accessing UMLS concepts have been made easy through the *MetaMap-2010* which is an open source toolkit from the National Library of Medicine (NLM), explained in details in [3]. Metamap is capable of parsing and finding medical concepts with options allowing users to disambiguate and filter UMLS concepts in the text. We use metamap in two ways, firstly to identify keywords from a verbose query and secondly to extract candidate terms for query expansion. Figure 1 demonstrates a comparison of keyword search, expansion using the best sematic type and a plain run, confirming no improvement in using only keyword candidates from the Metamap.

We used MetaMap to find trigger words pertaining to treatment or interventions in the queries, however, for document level processing, section headings were better clues in finding treatment or intervention related terms, this is explained in more detailed in section 6.

Each concept extracted from the metamap is associated to a semantic type, we parse all the test queries and extract 61 semantic labels and use them to filter candidate terms for expansion.

Figure 2 demonstrate the effect of 20 semantic types for expansion and a list of all the semantic types numbered from 1 to 21 are presented in Table 1. The process of expanding a query term using its semantic type is illustrated by an example as follow:
Table 2 is an output generated by Metamap and the terms appearing in brackets refer to semantic

[2]DBpedia:`http://dbpedia.org/About`

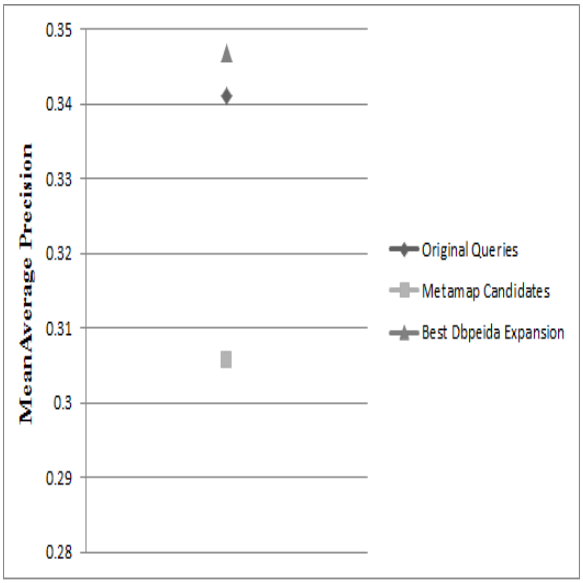


Figure 1: Keyword search using metamap in comparison with semantic expansion and plain topics

types associated to each concept. Identified concepts were later expanded by DBpedia to form a verbose query. Finally we add the expansion terms to the original query and repeat the same for each semantic type.

| Number | Semantic Type |
|---|---|
| 1 | Body Substance |
| 2 | Organic Chemical,Pharmacologic Substance |
| 3 | Health Care Related Organization,Manufactured Object |
| 4 | Educational Activity |
| 5 | Professional or Occupational Group |
| 6 | Age Group |
| 7 | Eukaryote |
| 8 | Family Group |
| 9 | Pharmacologic Substance |
| 10 | Clinical Attribute |
| 11 | Neoplastic Process |
| 12 | Health Care Activity |
| 13 | Manufactured Object |
| 14 | Finding |
| 15 | Pathologic Function |
| 16 | Spatial Concept |
| 17 | Patient or Disabled Group |
| 18 | Medical Device |
| 19 | Amino Acid, Peptide, or Protein,Biologically Active Substance |
| 20 | Disease or Syndrome |
| 21 | Mental or Behavioral Dysfunction |

Table 1: 21 semantic types from the output of Metamap

Figure 3 demonstrates our early experiment on 6 different methods for expansion over the 4 sample topics. Although removing negated terms are not beneficial for the given 4 sample queries we anticipated that test queries would possibly have multiple negated questions and it is vital

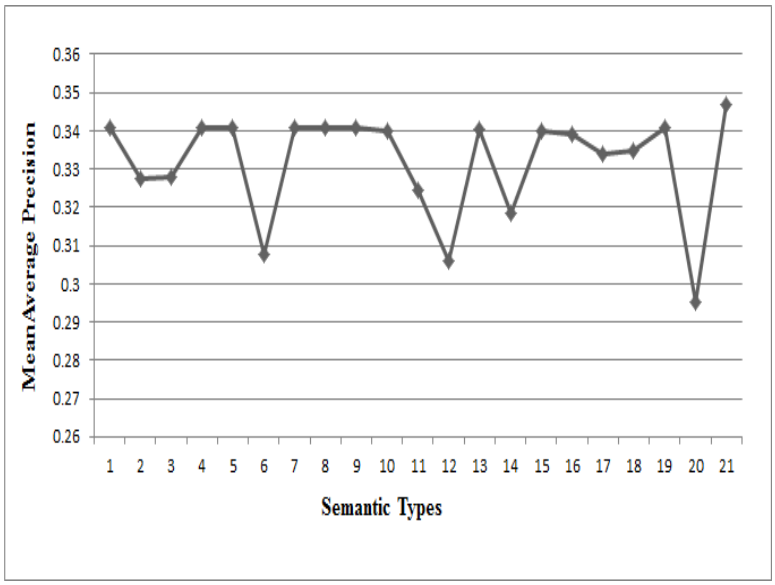


Figure 2: Expansion using 21 different semantic types from Metamap with TF.IDF

| | |
|---|---|
| Phrase: "with hearing loss" | |
| Meta Candidates (6): | |
| 1000 C1384666:Hearing Loss | (hearing impairment) [Finding] |
| 1000 C2029884:hearing loss | (hearing loss by exam) [Finding] |
| 861 C0018767:Hearing | [Physiologic Function] |
| 861 C1455844:Hearing | (Hearing examination finding) [Finding] |
| 861 C1517945:Loss | [Quantitative Concept] |
| 861 C2015933:hearing | (outcomes otolaryngology hearing) [Finding] |
| Meta Mapping (1000): | |
| 1000 C1384666:Hearing Loss | (hearing impairment) [Finding] |
| Meta Mapping (1000): | |
| 1000 C2029884:hearing loss | (hearing loss by exam) [Finding] |

Table 2: Output generated by metamap

to remove negated terms from all the queries. However, later we perform an alternative to this approach by replacing negated terms in the documents, this is explained in more detail in the next section.

We install three different ranking algorithms to find a consistent trend on the performance of various methods. Numbers along the x-axis denote the expansion method described in Table 3.

| Expansion | Description |
|---|---|
| 1 | No expansion |
| 2 | Remove negated terms |
| 3 | Remove negated and Wikipedia expansion |
| 4 | Remove negated and UMLS expansion |
| 5 | Remove negated and MeSH expansion |
| 6 | Remove negated and combined expansion |

Table 3: Description of expansion methods

Removing negated terms and expanding queries using Wikipedia consistently obtains better results, note that the low average precision of the PL2 model demonstrated in Figure 3 is mainly due to the higher number of unjudged documents for this particular system while all the top 20

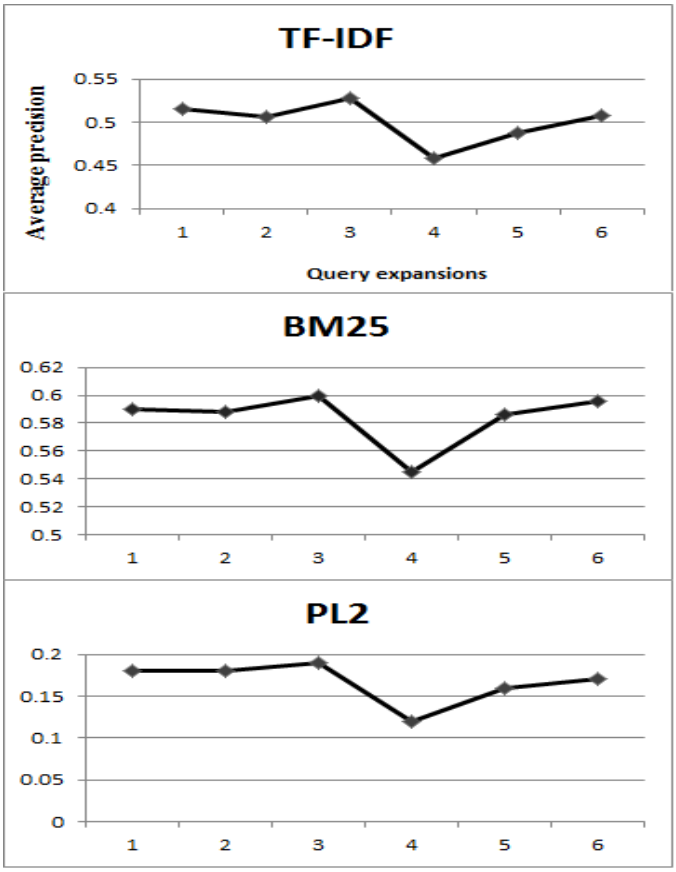


Figure 3: Performance of 3 systems on 6 different types of expansion described in Table 3

documents retrieved for the other 2 systems were judged.

Document expansion was another intuitive aspect that we and other groups [6, 10, 7] looked at, particularly, two sections of the reports were likely to benefit from expansion, namely: "Discharge Diagnosis" and "Admit Diagnosis". They contain ICD codes which were not mentioned in any of the sample queries, indicating the need for expansion. We use a full description of ICD9 table to expand these codes and to index them. Extending documents, through mapping the ICD codes, led to a relative improvement of the map score of the plain run, increasing from 0.3410 to 0.3570, however this improvement was not significant.

# 5 Negation

NegEx [5] algorithm is mostly known to Text Mining researchers for finding terms used in negative senses. This is crucial in retrieving a medical document as practitioners often report the absence of a disease or symptoms, which can potentially mislead any ranking algorithm. If a query targets a population without a particular disease $D$ or treatment $T$ for instance, search engines have no way of penalizing documents for containing the word $D$ or $T$ by default, therefore we identify and remove the negated terms from the queries initially and later annotate the documents negating the triggered terms.

An example of a query with negated term is as follows:
Query: patients with a BMI $>$ 40 without hypertension. Another alternative that we tested later was to modify the occurrence of negated nouns with a prefix "no". So the terms *without hypertension* would be replaced by : "nohypertension". However if the negated terms occur in the

| |
|---|
| PRESENT_HISTORY:hearing loss DIS_DIAGNOSIS:hearing loss ELSE:with hearing loss |
| PRESENT_HISTORY:complicated gerd DIS_DIAGNOSIS:complicated gerd ELSE:who receive endoscopy |

Table 4: Fielding 2 test topics

| VisitBased | | | ReportBased | | |
|---|---|---|---|---|---|
| map | bpref | p_10 | map | bpref | p_10 |
| 0.35 | 0.47 | 0.47 | 0.34 | 0.44 | 0.46 |

Table 5: Comparison of Visit and Report based Indices

queries it has a risk of underestimating those relevant documents in which "hypertension" is not mentioned. Our experiment with negation on a plain run, doesn't gain in any improvement.

Other participants of the Medical Trec conference have done different experiments with Negation. University of Galsgow [10] introduce a new tokeniser (called NegExTokeniser) and Cengage Learning [8] additionally uses their own algorithm to find uncertainty of a concept. Both of the groups incorporate information from the negation and report margninal improvements.

# 6 Query Transformation and Field search

There are a few ways to make the language of query and document match through expanding the document, query or both. We defined a fixed array of values for each mention of Elderly, Adult, Young, Female and male in the queries.

We divide each document into 9 sections to perform fielded search, assuming that queries contain parts relevant to varying sections in the documents. These fields were identified using regular expression and separated using end of the section patterns.

We utilized field search by segmenting queries into fields such as converting age related terms, like elderly or young to the style used in the documents, and words followed by 'taking' or 'patients who are on' were grouped under the Medication field. Table 4 is an example of a conversion of the first two test queries into their corresponding fielded topics. However, Terrier by default imposes a boolean AND on the terms defined under fields; this experiment did not show any improvement over the plain run, as the number of retrieved documents significantly dropped resulted by aggressive AND restrictions. Another alternative to using fields is Per Field Normalization, Macdonald et al. [11] report great improvement in using it against the WebCLEF collection. Query terms are normalized per field assigning different weights to each, and the logic behind this technique is that varying term distribution and importance of each field such as title and content of web pages implies weighting each field differently.

# 7 Indexing and Retrieval

We utilized Terrier 3.5 open source search engine [1] for indexing purpose, and rely on the Terrier's default stemmer and stop word list for queries and documents. We also remove all the terms with low idf such as patient which has a very high term frequency in the collection.

TREC participants were provided with a mapping table having to retrieve a visit for each topic rather than a report file which was just a part of a complete visit. We indexed the collection initially by treating each report as a document and later an aggregate of reports as a single visit. The results show that the latter way is more effective and obtains a higher score as shown in table 5, however, the t-test does not show any significant difference.

For retrieval we mainly tested 3 different ranking algorithms using Terrier search engine. Before the release of test queries we tested our systems to configure our final retrieval systems, however,

the lack of relevance judgements prevented us from making a confirmed decision. After manually doing more relevance judgement we found that BM25 with default parameters do not perform as good as other 2 systems, this was confirmed by setting BM25 for the Lucene search engine which resulted in low performance again. Parameters used for the Okapi 25 are given in the equation 1.

$$BM25(Q,D) = \sum_{(t \in Q)} w_t . \frac{(k_1+1) f_{d,t}}{K + f_{d,t}} . \frac{(k_3+1) f_{q,t}}{k_3 + f_{q,t}} \quad (1)$$

where $K = k_1.((1-b) + \frac{b.L_d}{AL})$
$Q$ query
$D$ document
$w_t$ is the Robertson-Spark Jones weight
$f_{d,t}$ and $f_{q,t}$ are the number of occurrences of term t in the document and query, respectively
$L_d$ and $AL$ are the document length and average document length
$k_1$,$k_3$ and b are parameters that are determined empirically
$k_1 = 1.2$, b = 0.75 and $k_3 = 8$

$k_3$ is often set to 0 because of the short nature of queries which often implies that words in the queries do not occur more than once.

We believe that a better choice of parameters for Okapi can lead to improved results.

Other two ranking metrics used were PL2 – an advanced model from Divergence-from-Randomness, and the Lemur version of TF.IDF. For more details about the PL2 ranking algorithm please refer to [2].

# 8 Analysis of Results

Arguably existence of significance difference is more probable with large collections, however, recently there is more emphasis on performing tests that demonstrate the existence of any significance difference. Although our final runs were a blend of multiple experiments we intended to find the main factor, contributing to the goodness or weakness of each run, and hence we decompose each step and report significance test. We perform student t-test across 4 aspects of the experiments, explained in this section.

Firstly per visit indexing gained in higher scores, but the t-test does not show any statistical difference, with the p value $> 0.9$, RMIT runs, albeit, perform per report indexing. Indexing visits using SinglePass indexer from the TERRIER toolkit, however, was more efficient in terms of the speed.

Choosing the ranking algorithm involved testing 3 different ranking metrics, of which BM25 performs significantly worse than other two metrics: PL2 and Lemur version of TF.IDF with p value equal to 0. Default parameters of BM25 is given in section 7, weak scores from BM25 was quite questionable and made us to try it on the Lucene search engine, however, the low performance of BM25 was rather consistent. Automatically setting the values of BM25 parameters was however, out of the scope of this paper.

The remaining tests proved the absence of significance difference on using the best semantic type expansion and extension of ICD codes in the documents versus a plain run.

# 9 RMIT Runs

The 4 runs mainly differ in the combination of query expansion and translation used. Automatic query processing and translation was used to expand particular words that need expansion to match the language of clinical records, Table 6 demonstrates a short description for each individual run.

| Run Identifier | Description |
|---|---|
| Run-1 | Query Translation + Query Expansion + remove negated + Stemmed |
| Run-2 | Query Expansion + Stemmed |
| Run-3 | Query Translation + Stemmed |
| Run-4 | Similar to Run-1 using Lemur version of TF-IDF |

Table 6: Our 4 distinct runs in a nutshell

# 10 Evaluation

Prior to the release of result from the TREC, in addition to using extended judgments we utilized the automatic judgment method proposed in [9]. Trec evaluation result proves consistency to our automatic evaluation, we learn that the best ranking algorithm is the Lemur version of TF.IDF followed by PL2.

Table 7 shows the final result for our four distinct runs, three of which contributed to the pool and one that remained unjudged. However, NIST looked at retrieved documents for unjudged runs at various cut-offs and reports that the coverage is fairly reasonable.

| Metric | Run1 | Run2 | Run3 | Run4 |
|---|---|---|---|---|
| MAP | 0.3425 | 0.3240 | 0.3038 | 0.3527 |
| P@10 | 0.5235 | 0.5088 | 0.4265 | 0.4882 |
| B-pref | 0.4583 | 0.4397 | 0.4509 | 0.4580 |

Table 7: Evaluation of 4 distinct runs

# 11 Conclusion

Our runs perform above the median of the 47 judged and 80 unjudged submissions. All runs perform relatively equal with minimal differences, Table 8 compares our 4th run against best and median of all the runs. Combination of query transformation and expansion using combination of external sources while using the Lemur version of TF.IDF for ranking documents, seem to yield the best outcome.

# 12 Acknowledgements

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

| | | Judged | | Unjudged | |
|---|---|---|---|---|---|
| Metrics | RMIT-Run4 | Best | Median | Best | Median |
| P@10 | 0.4882 | 0.8765 | 0.4765 | 0.8588 | 0.4441 |
| B-pref | 0.4580 | 0.7607 | 4115 | 0.7577 | 0.4340 |
| R-prec | 0.3653 | 0.6095 | 0.3087 | 0.5983 | 0.3047 |

Table 8: Comparison of RMIT's 4th run against median and best runs over all judged and unjudged submissions

# References


[1] Terrier search engine. `http://terrier.org/`.

[2] G. Amati and C.J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389, 2002.

[3] 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.

[4] E.V. Bernstam, J.R. Herskovic, Y. Aphinyanaphongs, C.F. Aliferis, M.G. Sriram, and W.R. Hersh. Using citation data to improve retrieval from medline. *Journal of the American Medical Informatics Association*, 13(1):96–105, 2006.

[5] W.W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, and B.G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001.

[6] Antonio JImerno-Yepes Russell Loane Bastien Rance Francois Lang Nicholas Ide Emilia Apostolova Alan R. Aronson Dina Demner Fushman, Swapna Abhyankar. A knowledge based approach to medical records retrieval. In *Proceedings of TREC*, 2011.

[7] S. Karimi, D. Martinez, S. Ghodke, L. Zhang, H. Suominen, and L. Cavedon. Search for medical records: Nicta at trec 2011 medical track. 2011.

[8] B. King, L. Wang, and I. Provalov. Cengage learning at trec 2011 medical track. In *Proceedings of TREC*, 2011.

[9] B. Koopman, P. Bruza, L. Sitbon, and M. Lawley. Evaluating medical information retrieval. In *ACM SIGIR Forum*. ACM, 2011.

[10] N. Limsopatham, C. Macdonald, I. Ounis, G. McDonald, and M.M. Bouamrane. University of glasgow at medical records track: Experiments with terrier. In *Proceedings of TREC*, 2011.

[11] C. Macdonald, V. Plachouras, B. He, C. Lioma, and I. Ounis. University of glasgow at webclef 2005: Experiments in per-field normalisation and language specific stemming. *Accessing Multilingual Information Repositories*, pages 898–907, 2006.