NICTA and UBC at the TREC 2012 Medical Track

David Martinez[†] Arantxa Otegi* Eneko Agirre*
[†]NICTA and CIS Department, University of Melbourne, Australia
*IXA NLP Group, University of the Basque Country, Donostia, Basque Country

Abstract

We introduce two heterogeneous query expansion techniques, and a combined system to the TREC 2012 Medical Track. Our methods are based on external resources that provide expansion concepts related to the query terms, by means of the PageRank algorithm, and simple rules based on UMLS Semantic Types. In this paper we show that our systems are able to reach competitive performances at both the TREC-2011 and TREC-2012 tasks.

1 Introduction

In this paper we present the combined submission of the teams NICTA and UBC, which focuses on query expansion techniques. For this edition we build upon the NICTA-2011 systems [8], and we incorporate the Personalised PageRank algorithm in order to select the most similar concepts to the query terms, and then use them for query expansion.

The NICTA system was ranked 6th on the 2011 Medical Track (with regards to the Bpref measure). We did minimal changes to this knowledge-based query expansion method, and centered our efforts in combining this technique with a graph-based expansion approach from the UBC team, namely Personalised PageRank.

Personalized PageRank [6] has been successfully used in Natural Language Processing tasks such as Word Sense Disambiguation [3, 4, 5, 11] and word similarity [1, 2]. It has been applied both to a general purpose lexical knowledge-base such as WordNet [1, 2, 3, 4] and also to UMLS [5, 11]. In addition, recent results show that it is useful to improve ad-hoc IR with WordNet [9]. In this work, we apply it on UMLS in order to improve results over the TREC task.

Our final scores show that query expansion is beneficial over the baseline methods; specially over the TREC-2011 queries, where it reaches the performance of the best 2011 systems. For the TREC-2012 query-set the results were far from the best performing system, but above the median of the submissions.

2 Method

We present here the steps of our approach: we start by describing how we processed the TREC document collection; then we explain our query processing method, including the expansion techniques; finally we detail our indexing and searching approaches.

Field	Description
ADMITDIAG	diagnostics during admission
AGE	patients age by decades, for example age30 means people in their thirties
ALLERGIES	allergies listed in the report
CHIEFCOMP	chief complaint, this may be equal to diagnostics during admission
DISCHDIAG	discharge diagnostics
GENDER	patient's gender extracted from text and represented as gendermale and genderfemale
HISTORY	history of the patient's medical condition or past medical illness
MEDICATIONS	medications
PRESTHIS	present illness medical history
PASTHIS	past medical history
REPORT	all the free text information, including history, past and present, and allergies

Table 1: List of fields defined for Boolean search.

2.1 Processing the Document Collection

We apply the same pipeline as in [8] for processing the document collection. We start by expanding the mentions of ICD9 codes[1] of admission and discharge diagnoses in the metadata with their text descriptions. Both the original code and expanded forms were included for indexing.

The documents contain different sections, with their corresponding headings. We rely on hand-crafted pattern-matching rules to identify the main headings, in order to build different indices and allow for field-based search. The list of the fields we cover is given in Table 1. Apart from these fields, we built rules to identify and normalise some demographic information, such as gender, age, and other specific conditions (such as weight) mentioned in the text.

We also ran *NegEx*[2] over the entire collection in order to detect negated phrases. We rely on the in-built Negex parser of *MetaMap-2010*, which specifies which of the identified phrases appear to be negated. We use this information to build a separate index that converts negated terms that are majority in a given document, into a new representation, where the negated phrase is transformed into a single word, with no space, and with a "*no*" prefix: e.g., if negation is implied for "chronic back pain", all instances of "chronic back pain" are replaced with the word "nochronicbackpain". Our aim with this index is to avoid matching cases where the term appears negated in the document more often than as positive.

Finally, we indexed the collection with and without the Porter stemmer.

2.2 Processing Queries

We describe first our methods to identify fields in the query, and then our different expansion approaches.

2.2.1 Identifying Fields in the Query

We developed a set of manually constructed patterns to map query terms into the available fields (Table 1). These patterns — formed based on the sample clinical questions provided by the National Library of Medicine (NLM) [7] — covered seven broad categories of age, weight (using body mass index), diagnostics, treatments, medications, history, allergies, and

[1] International Statistical Classification of Diseases and Related Health Problems: `http://en.wikipedia.org/wiki/List_of_ICD-9_codes`

[2] `http://code.google.com/p/negex/`

What	Pattern	Translation
Gender	women/female	GENDER:gendermale
	men/male	GENDER:gendermale
Age	young adult	AGE:(age20 age30 age40)
	younger/young	AGE:(agebirth12 ageteen age20 age30 age40)
	adult	AGE:(age20 age30 age40 age50 age60 age70 age80 age90)
Weight		
Treatments	taking X (who|with|without|treated)	MEDICATIONS:X
	who are on X	MEDICATIONS:X
	patients on X for Y	MEDICATIONS:X
Admission	admitted (for|with) X who	CHIEFCOMP:X OR ADMITDIAG:X
Diagnostics	treated for X (who|during|while)	PRESTHIS:X OR DISCHDIAG:X
	(patients with|men with|women with) X	PRESTHIS:X OR DISCHDIAG:X
	who were discharged X	DISCHDIAG:X
History	with a* history of X (who|now)	HISTORY:X
Allergy	with X allergy	ALLERGY:X
	without allergy	ALLERGY:(noallergies)
Abbreviation	seen in the er|presented to the er	REPORT:("emergency room" OR ER)

Table 2: Rules (patterns in the queries and their translations) used in the query transformation step. Words that are all in capital letters are field names.

abbreviations. For example, if a query contained "elderly patients", we expanded "elderly" with an equivalent age field that covered people in their 60s to 90+. Table 2 shows the details of the selected transformation rules. For example the query:

Elderly patients with ventilator-associated pneumonia

is translated to:

> *PRESTHIS:(ventilator associated pneumonia) OR DISCHDIAG:(ventilator associated pneumonia) OR AGE:(age60 age70 age80 age90) OR REPORT:(elderly with ventilator associated pneumonia).*

A small number of abbreviations, such as ER (emergency room), were also expanded in the queries.

2.2.2 Query Expansion using Semantic Types (ST)

We leveraged external resources to add new terms to our queries, by identifying terms that are strongly related to the query terms. Specifically, we focused on query terms that represent medical categorical concepts (e.g. disease categories). For example, for the query below, we added terms falling under the category of "atypical antipsychotics":

Patients taking atypical antipsychotics

Our approach to expansion used two main knowledge sources: the UMLS Metathesaurus (version 2010AA) and DBpedia. In order to select expansion candidates we used *MetaMap-2010* from the National Library of Medicine (NLM). We defined manual expansion rules from these resources based on the sample queries and 50 priority queries from the NLM priority list.

For our final expansion system, we first applied MetaMap to identify phrases linked to terms in the UMLS Metathesaurus. The matched concepts were then used as candidate terms to be expanded; in some cases terms consisted of a primary term followed by a parenthesized description — such as "Intervention (Surgical and medical procedures)" — and in such cases we treated them as separate candidate terms.

Each candidate term had a Semantic Type (ST) associated with it in the MetaMap output. We used STs to define two expansion groups: safe expansion (for terms which STs include the string "Pharmacologic Substance") and filtered expansion (for terms whose ST is "Therapeutic or Preventive Procedure"). Candidate terms that did not belong to these groups were discarded. For the rest, if they were listed as "category" in DBpedia[3], we extracted all of the terms listed under the category as our expansion terms. For "safe expansion" the output was the full list of expansion terms; for "filtered expansion", we removed terms which are not UMLS concepts by applying MetaMap to each term.

In our implementation, we defined a small set of stop-categories that would have otherwise produced undesirable expansions. The following terms were excluded from expansion: "administration", "AMA", "diagnosis", "drug", "functional concept", "medication", and "surgery". We also removed terms with the following strings from the DBpedia output: "code", "history", "mechanism", "poisoning", "toxicity", and "withdrawal".

During the development process, we also explored expansion using hierarchical relations from the UMLS Metathesaurus, by selecting all the terms in the hyponym concepts; however, we observed that DBpedia offered a higher coverage of some domains, such as newly developed drugs, and it also showed less risk of over-expansion. For instance, one sample query contained the term "atypical antipshychotic", which UMLS expanded with 8 more specific drugs (e.g. "Clozapine"). DBpedia, however, identified the same set of drugs, as well as a further 22 new drug and brand names, which seemed correct after manual analysis, and had a stronger presence in the collection.

2.2.3 Query Expansion using Personalised PageRank

For this approach, we use a graph algorithm based on random walks over the graph representation of a knowledge-base of concepts and relations, to obtain concepts related to the queries. The UMLS Metathesaurus is used as the knowledge-base, and we represent UMLS as a graph.

Apart from concepts, UMLS Metathesaurus also contains a wide range of information about the relations between concepts in the form of database tables. The MRREL table lists relations between concepts like "parent", "can be qualified by" or "related and possibly synonymous" among others. The MRCOC table contains co-occurrence relations between concepts, that is, relations between similar concepts or different concepts that share an important connection. In order to obtain the graph structure of UMLS, we simply treat the concepts in UMLS as vertices, and the relations listed in the MRREL and MRCOC tables as edges. No weights are used for the relations that are extracted from the MRREL table.

Given a query and the graph-based representation of UMLS, we obtain a ranked list of related concepts as follows:

1. We first run MetaMap and identify the UMLS concepts in the query, we explore two variants: with and without the in-built Word Sense Disambiguation (WSD) module.

[3] http://wiki.dbpedia.org/OnlineAccess

We also rely on the Negex module to remove negated concepts. Note that in cases where we rely on field search, we treat each field as a separate query for this kind of expansion.

2. We then assign a uniform probability distribution to the concepts found in the query. The rest of nodes are initialized to zero.

3. We compute personalized PageRank [6] over the graph, using the previous distribution as the initial distribution, and we produce a probability distribution over UMLS concepts. The higher the probability for a concept, the more related it is to the given text.

Basically, personalized PageRank is computed by modifying the random jump distribution vector in the traditional PageRank equation. In our case, we concentrate all probability mass in the concepts identified in the query.

Let G be a graph with N vertices $v_1, \ldots, v_N$ and d_i be the outdegree of node i; let M be a $N \times N$ transition probability matrix, where $M_{ji} = \frac{1}{d_i}$ if a link from i to j exists, and zero otherwise. Then, the calculation of the *PageRank vector* $\mathbf{Pr}$ over G is equivalent to resolving Equation (1).

$$\mathbf{Pr} = cM\mathbf{Pr} + (1 - c)\mathbf{v} \tag{1}$$

In the equation, $\mathbf{v}$ is a $N \times 1$ vector and c is the so-called *damping factor*, a scalar value between 0 and 1. The first term of the sum on the equation models the voting scheme described in the beginning of the section. The second term represents, loosely speaking, the probability of a surfer randomly jumping to any node, e.g. without following any paths on the graph. The damping factor, usually set in the [0.85..0.95] range, models the way in which these two terms are combined at each step.

The second term on Eq. (1) can also be seen as a smoothing factor that makes any graph fulfill the property of being aperiodic and irreducible, and thus guarantees that the PageRank calculation converges to a unique stationary distribution.

In the traditional PageRank formulation the vector $\mathbf{v}$ is a stochastic normalized vector whose element values are all $\frac{1}{N}$, thus assigning equal probabilities to all nodes in the graph in case of random jumps. In the case of personalized PageRank as used here, $\mathbf{v}$ is initialized with uniform probabilities for the concepts in the query, and 0 for the rest of terms.

PageRank is actually calculated by applying an iterative algorithm which computes Eq. (1) successively until a fixed number of iterations are executed. In our case, we used a publicly available implementation[4], with the default values provided by the software, i.e. a damping value of 0.95, and 30 iterations.

In order to select the expansion terms from the ranking of concepts, we use a threshold value to retrieve the top concepts, and then we obtain all the terms that appear under each concept in the UMLS Metathesaurus. We explored two approaches to determine the cut-off value: (i) select the top k concepts, or (ii) select all the concepts with weights above a given t threshold. Our preliminary experiments over the TREC-2011 dataset suggested that the former approach was able to provide better performances for different settings, and we decided to use the top k concepts for our experiments.

[4] http://ixa2.si.ehu.es/ukb/

2.2.4 Combined Query Expansion

In order to combine our two different expansion techniques, we can simply merge the terms from each expansion source into a joint query. Another approach that we explored is to rely on the expanded terms from the ST-expansion to initialise the PageRank method. We report the results of the two methods in our experiments.

2.3 Indexing and Searching

We first distinguish between two types of indexing in our runs: *visit-based* and *report-based*. In the former approach, all related reports for a visit were concatenated (removing duplicate diagnostics codes) to create a single "multi-document" item for indexing. We refer to the former approach as VISIT, and as REPORT to the latter.

As explained in Section 2.1, we also generate different indexes depending on the use of Negex or not (NEG/NONEG), the use of separate fields or not (FIELDS/COMBINED), or the application of stemming (STEM/NOSTEM). When we rely on field search, a Boolean search over the fields is followed by ranking.

We used stop-word removal both in query processing and indexing; however, we augmented the typical list of stop-words with *patient*, and removed single characters, *and*, *or*, *not*, and *no* from the list.

The search engine used for indexing and searching in our runs was Apache Lucene (v3.2); we used both the BM25 and *tf-idf* ranking algorithms for Lucene [10].

3 Results over the TREC-2011 query set

We first tested different combinations of our main approaches over the TREC-2011 query set and collection, in order to select the most promising configurations for TREC-2012. We relied on the same evaluation metric that was used in TREC-2011: Bpref.

We performed three main experiments:

- PageRank without Semantic Type (ST) expansion
- Combine ST expansion and PageRank, without field indexing
- Combine ST expansion and PageRank, with field indexing

As mentioned above, when we combine PageRank and ST, we have to choose if we want to apply PageRank over the query concepts, or over the ST-expanded concept set. We present the results for the two different settings in most of our experiments. There are other two alternatives when applying PageRank: to perform WSD prior to choosing the initial concepts, or not to use WSD. We report here the results of the two variants. Finally, we also need to set a threshold to decide the number of top concepts to use. We performed preliminary experiments using two types of thresholds: weight-based (i.e. choose all the concepts above the cut-off PageRank weight) and ranking-based (i.e. select all the concepts in the top k positions). Our initial experiments showed better performance with the ranking-based threshold, and we used this method for our main experiments. We report the results for the best and worst cut-offs in the range 3-20 over the TREC-2011 dataset.

We start our analysis by evaluating the performance of PageRank without ST expansions. In this case we also need to decide whether we parse the query before applying

System	WSD	Best Bpref (threshold)	Worst Bpref (threshold)
Base system	No	0.5218	**0.5218**
PageRank first	No	**0.5438** (3)	0.5203 (18)
PageRank first	Yes	0.5373 (9)	0.5026 (3)
Query Transformation first	No	0.5427(15)	0.3719 (3)
Query Transformation first	Yes	0.5412 (8)	0.4048 (3)

Table 3: Performance of different PageRank settings over the TREC-2011 query set, together with the baseline. Best results per column in bold.

System	WSD	Best Bpref (threshold)	Worst Bpref (threshold)
Base system	No	0.5078	0.5078
PageRank first	No	**0.5655** (3)	**0.5293** (18)
PageRank first	Yes	0.5488 (3)	0.5277 (20)
ST Expansion first	No	0.5501(9)	0.4923 (3)
ST Expansion first	Yes	0.5480 (5)	0.4997 (3)

Table 4: Results combining PageRank and ST expansion (NEG+VISIT+STEM+COMBINED and TF-IDF), the best results per column are given in bold.

PageRank or not. The results over the TREC-2011 query set are given in Table 3, together with the basic system without PageRank. For this experiment we chose the index NEG+VISIT+STEM+COMBINED and TF-IDF ranking as basic system, since it achieved the highest performance in previous experiments when no ST expansions were used.

We can see that the system achieves its best performances when applying PageRank first, and that we are able to improve over the baseline. WSD does not seem to be helpful, and starting with all the concepts from MetaMap (not only the disambiguated ones) is the best strategy for this experiment.

For our next experiment we combine the ST expansion with PageRank. As base configuration, we rely on the same index and ranking used in the previous test (NEG+VISIT+STEM+COMBINED and TF-IDF). The results of this experiment are given in Table 4.

There is a larger improvement over the baseline in this case, and even the worst thresholds improve the baseline when PageRank is applied first. Note that the best results are similar to the best official submission for the TREC 2011 challenge. Again, the best performance is achieved without WSD.

Next we performed a similar experiment by using report indexing (instead of visits), and no stemming; we chose this indexing because it was also competitive, and we observed clear differences over the outputs of these settings in previous experiments. We present the results of this experiment in Table 5.

These results reach the highest Bpref score so far, and are more robust regarding the lower bounds. Again, the best strategy is to apply PageRank first, and not to use WSD in the process.

We also explored the use of fields in the indexing. This approach obtained worse perfor-

System	WSD	Best Bpref (threshold)	Worst Bpref (threshold)
Base system	No	0.4895	0.4895
PageRank first	No	**0.5789** (3)	**0.5422** (10)
PageRank first	Yes	0.5495 (3)	0.5226 (10)
ST Expansion first	No	0.5642 (7)	0.5008 (4)
ST Expansion first	Yes	0.5468 (5)	0.5041 (3)

Table 5: Results combining PageRank and ST expansion (NEG+REPORT+NOSTEM+COMBINED and TF-IDF), the best results per column are given in bold.

System	WSD	Best Bpref (threshold)	Worst Bpref (threshold)
Base system	No	0.4802	**0.4802**
PageRank first	No	**0.5127** (7)	0.4561 (19)
PageRank first	Yes	0.4955 (7)	0.4540 (19)

Table 6: Results combining PageRank and ST expansion using fields, the best result per column are given in bold.

mance that combining fields in our previous experiments, and we only performed two runs, always applying PageRank first. The results are shown in Table 6. We can see that the gains are smaller than in previous configurations, and there is a big drop in the case of the worst threshold.

4 Results over the TREC-2012 query set

In order to diversify, we chose four configurations that achieved good performance over the TREC-2011 dataset. For all our runs, we relied on the COMBINED index and we did not process negations for the documents (only for the queries), we also use TF-IDF in all the submitted runs:

- NICTAUBC1: Combined expansion, PageRank first (threshold = 3), index REPORT+STM
- NICTAUBC2: Combined expansion, ST expansion first (threshold = 4), index REPORT+NOSTM
- NICTAUBC4: ST expansion, index VISIT+STM
- NICTAUBC6: Combined expansion, ST expansion first (threshold = 6), index VISIT+NOSTM

The results of the different systems are given in Table 7, together with the median and best results for the automatic runs. We can see that NICTAUBC4 is our best performing system, scoring well above the median for all metrics. This means that we obtain the best performance when we only rely on ST expansion, and unlike our TREC-2011 experiments, adding PageRank is not helpful for these runs over the 2012 query set.

System	infAP	infNDCG	R-prec	P@10
NICTAUBC1	0.1947	0.4362	0.3053	0.3915
NICTAUBC2	0.1912	0.4450	0.3023	0.4362
NICTAUBC4	**0.2162**	**0.4870**	**0.3417**	**0.5170**
NICTAUBC6	0.1837	0.4193	0.2950	0.4170
Best Automatic	0.4238	0.7461	0.5428	0.8149
Median Automatic	0.1695	0.4243	0.2935	0.4702

Table 7: Official results for TREC-2012, our best performances are given in bold.

Expansion	Configuration	infAP	infNDCG
No expansion	REPORT+STEM+FIELDS and BM25	0.1793	0.4168
PageRank	REPORT+STEM+COMBINED (thr=4) and TF-IDF	0.2176	0.4704
ST	VISIT+STEM+COMBINED and TF-IDF	0.2162	0.4870
Combined	VISIT+STEM+COMBINED (thr=4) and TF-IDF	0.2252	0.4790

Table 8: Best configuration for the types of systems we developed.

After the qrels were released, we tested the performances of our different systems, and we show in Table 8 the performances of the best configurations of the expansions we developed. Again, we can see that the ST approach performs best, with the best result being the one that we submitted. For the combined system, we can see that we can see that the setting of the threshold for PageRank is an important issue to be tacked.

5 Conclusions

Our expansion techniques showed a different behaviour over the TREC-2011 and TREC-2012 query sets, with the promising initial results of PageRank not translating so well into this year's challenge. We plan to perform a thorough analysis of the different queries, in order to learn the reasons of this, and explore better ways to develop expansion techniques that benefit from the combined expansion approach over medical data.

Acknowledgments

NICTA is funded by the Australian government as represented by Department of Broadband, Communication and Digital Economy, and the Australian Research Council through the ICT Centre of Excellence programme. This research was partially funded by the Ministry of Economy under grant TIN2009-14715-C04-01 (KNOW2 project).

References

[1] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North*

American Chapter of the Association for Computational Linguistics, NAACL '09, pages 19–27. Association for Computational Linguistics, 2009.

[2] E. Agirre, M. Cuadros, G. Rigau, and A. Soroa. Exploring knowledge bases for similarity. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010. European Language Resources Association (ELRA).

[3] E. Agirre, O. L. de Lacalle, and A. Soroa. Knowledge-Based WSD on Specific Domains: Performing better than Generic Supervised WSD. In *Proceedings of IJCAI*, pages 1501–1506, Pasadena, USA, 2009.

[4] E. Agirre and A. Soroa. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 33–41. Association for Computational Linguistics, 2009.

[5] E. Agirre, A. Soroa, and M. Stevenson. Graph-based word sense disambiguation of biomedical documents. *Bioinformatics*, 26(22):2889–2896, Nov. 2010.

[6] T. H. Haveliwala. Topic-sensitive PageRank. In *Proceedings of WWW '02*, pages 517–526, 2002.

[7] Institute of Medicine. 100 initial priority topics for comparative effectiveness research, 2009.

[8] S. Karimi, D. Martinez, S. Ghodke, L. Zhang, H. Suominen, and L. Cavedon. Search for Medical Records: NICTA at TREC 2011 Medical Track. In *Proceedings of the Text Retrieval Conference (TREC)*, 2012.

[9] A. Otegi, X. Arregi, and E. Agirre. Query expansion for ir using knowledge-based relatedness. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 1467–1471, Thailand, 2011.

[10] J. Perez-Iglesias, J. Perez-Aguera, V. Fresno, and Y. Feinstein. Integrating the probabilistic models BM25/BM25F into Lucene. *CoRR*, abs/0911.5046, 2009.

[11] M. Stevenson, E. Agirre, and A. Soroa. Exploiting domain information for word sense disambiguation of medical documents. *JAMIA*, 19(2):235–240, 2012.