NLM at TREC 2012 Medical Records track

Dina Demner-Fushman, Swapna Abhyankar, Antonio Jimeno-Yepes, Russell Loane, Francois Lang, James G. Mork, Nicholas Ide, Alan R. Aronson

National Library of Medicine, Bethesda, Maryland

Abstract

The NLM team used the relevance judgments for the 2011 Medical Records track (that focused on finding patients eligible for clinical studies) to analyze the components of our 2011 systems. The analysis showed that the components provided moderate improvements over the baseline (established submitting 2011 topics 'as is' to Lucene) for some topics and did not harm the results for any other topics. Our experiments confirmed that implementing methods (such as negation detection and section splitting) motivated by clinical text processing experience could improve identifying patients that meet complex criteria for inclusion in cohort studies. We therefore largely used the 2011 system with minor modifications for document processing.

We submitted three automatic runs: an Essie baseline run, and two Lucene runs that used the 2011 system with minor modifications. We also submitted an interactive run for which the queries were interactively modified using Essie until either the top ten retrieved documents appeared mostly relevant or no relevant documents could be found.

Our interactive queries submitted to Essie significantly outperformed all our other runs and were significantly above the medians for all submission types (achieving 0.37 infAP; 0.68 infNDCG; 0.75 P@10; and 0.48 R-prec). Interestingly, the values of the two metrics common for the two years of this track are very close to the values achieved in 2011. The hypothetical overall-best and best-manual performances are significantly better than our interactive run. Our Lucene run that used the topic frames and web-based expansion is significantly better than the Lucene baseline run and the medians (on all metrics but P@10 for the medians), but it is not significantly better than our other automatic runs. Our other automatic runs are not significantly above the medians. As in 2011, we conclude that the existing search engines are mature enough to support cohort selection tasks, and the quality of the queries could be significantly improved with a modest interactive effort.

1. Introduction

The 2012 TREC Medical Records track repeated the 2011 task and focused on finding patients who were eligible for inclusion in clinical studies. The track also reused the clinical narrative documents generated during the patients' hospital stays and collated into one visit for each hospital stay. A post-hoc analysis of our 2011 system using the 2011 Medical Records track relevance judgments suggested that our assumptions about preprocessing needed for clinical document retrieval were not likely to harm our performance. We therefore decided to use the 2011 systems with only minor modifications and bug fixes. The post-hoc experiments are discussed in Section 2.

Our efforts for the 2011 track document processing started with splitting documents into sections; then splitting each section into *Positive* (containing asserted findings, problems, and interventions), *Negative* (in which findings are negated), and *Speculative* (that includes all uncertain statements); identifying UMLS terms and expanding the recognized terms in the documents with their parents and children; translating the ICD-9 codes to their preferred terms in the UMLS; and extracting the patient's age and gender into structured fields. These document-preparation steps remained largely unchanged, except for the revision of the section splitting rules that is described in Section 3.

As in 2011, we indexed the documents using Essie and Lucene and translated the topics to question frames as described in our 2011 report (Demner-Fushman et al, 2011). We modified the frame-to-query translation rules to accommodate the revised document sections. The 2011 query expansion modules were also reused. Our experiments are discussed in Section 4.

We conclude the report with a preliminary analysis of our experiments and results and an in-depth discussion of the results of our interactive run.

2. Post-hoc analysis of the 2011 system

The post-hoc analysis focused on the following questions:

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2012		**00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
NLM at TREC 2012 Medical Records track	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **National Library of Medicine,8600 Rockville Pike,Bethesda,MD,20894**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

Presented at the Twenty-First Text REtrieval Conference (TREC 2012) held in Gaithersburg, Maryland, November 6-9, 2012. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA). U.S. Government or Federal Rights License

14. ABSTRACT

The NLM team used the relevance judgments for the 2011 Medical Records track (that focused on finding patients eligible for clinical studies) to analyze the components of our 2011 systems. The analysis showed that the components provided moderate improvements over the baseline (established submitting 2011 topics ?as is? to Lucene) for some topics and did not harm the results for any other topics. Our experiments confirmed that implementing methods (such as negation detection and section splitting) motivated by clinical text processing experience could improve identifying patients that meet complex criteria for inclusion in cohort studies. We therefore largely used the 2011 system with minor modifications for document processing. We submitted three automatic runs: an Essie baseline run, and two Lucene runs that used the 2011 system with minor modifications. We also submitted an interactive run for which the queries were interactively modified using Essie until either the top ten retrieved documents appeared mostly relevant or no relevant documents could be found. Our interactive queries submitted to Essie significantly outperformed all our other runs and were significantly above the medians for all submission types (achieving 0.37 infAP; 0.68 infNDCG; 0.75 P@10; and 0.48 R-prec). Interestingly, the values of the two metrics common for the two years of this track are very close to the values achieved in 2011. The hypothetical overall-best and best-manual performances are significantly better than our interactive run. Our Lucene run that used the topic frames and web-based expansion is significantly better than the Lucene baseline run and the medians (on all metrics but P@10 for the medians), but it is not significantly better than our other automatic runs. Our other automatic runs are not significantly above the medians. As in 2011, we conclude that the existing search engines are mature enough to support cohort selection tasks, and the quality of the queries could be significantly improved with a modest interactive effort.

15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**5**	

1) Was translating the original topics to structured frames useful?
2) Was segmenting the documents into sections and giving more weight to specific sections for a given frame slot useful?
3) Was giving more weight to positive text (all text that was not identified as negative, speculative, or in the family history section) useful?
4) Was query expansion useful?

To answer these questions we conducted experiments that compared the following conditions:

1) Searching the original text vs. searching positive fields.
2) Using pre-defined weights for the sections vs. using equal weights for all sections.
3) Using the original topics vs. the topic frames.
4) Query expansion vs. none (for Essie, rather than completely avoiding query expansion that could be achieved by requiring exact string match, we chose term expansion that allows term normalization to the base form in the Specialist Lexicon and might be viewed as an equivalent to stemming in Lucene.)

The Lucene experiments revealed a bug in our 2011 system – the positive text was limited to speculative and did not include the assertions. Once the bug was fixed, Lucene results were consistent with the Essie results shown in Table 1.

Table 1. Post-hoc evaluation of the 2011 system components

Run	P@10	Rprec	Bpref
Baseline (original topics, original documents)	0.4765	0.3438	0.7954
Topic frames, Original documents	0.5235	0.3699	0.8574
Topic frames, Document sections	0.4912	0.3366	0.8617
Topic frames, All positive text	0.5353	0.3718	0.8582
Topic frames, Positive text, Document sections	0.4941	0.3454	0.8581

The overall differences between the post-hoc runs were not significant. Inspecting the results for individual topics we decided that no individual component was consistently harming the system performance and each component could be useful for more complex queries.

3. Document segmentation

Segmenting clinical documents has to strike a balance between creating too many sections (which would increase the complexity of the queries) and failing to separate the sections bearing different types of information. Hoping to improve the impact of the section-base retrieval, we have revised our section splitting rules. The new sections are:

Preamble [the structured info at the top of each TREC document]
Addendum [extra information the clinician wants to make sure is in the record. Could be as short as details about a follow-up appointment, or could be the entire course of the hospitalization]
Admission_diagnosis [the diagnosis given by the clinician for why the patient was admitted]
Chief_complaint [the reason given by the patient for why the patient is there]
Final_diagnosis [the final diagnosis or list of diagnoses given by the clinician at the end of the admission (may not be the same as the admission diagnosis)]
Problem_list [unique to progress notes - a list of the patient's active problems during the hospitalization. Will likely overlap with past medical history, admission diagnosis and discharge diagnosis]
History_of_present_illness [summary of the patient's symptoms and other events before the patient was admitted to the hospital]
Past_medical_history [list of the patient's diagnoses before this hospitalization]
Family_history [list of the patient's relatives having the diseases for which the patient might have higher risks given the family history]
Social_history [patient's behavioral traits (such as smoking) and social circumstances that can influence the course of the hospital stay]
ROS [Review Of Systems -- a list of the patient's symptoms by organ system, both positive and negative. e.g., +cough, no wheezing; +nausea, no vomiting. Lots of overlap with history of present illness]
Home_meds [meds the patient was on before coming to the hospital]
Hospital_medications [meds the patient is on while in the hospital]
Discharge_meds [meds the patient is being discharged on]
Allergies [drug or other allergies]
Physical_exam [self-explanatory]
Lab_rad_results [results of lab and radiology tests]

Procedure_results [unique to progress notes, dc summaries, and er - give the results of surgical or other procedures that were done. Some overlap with lab_rad_results, but more specific to surgical or other types of procedures (e.g. cardiac cath, stress test)]
Consults [the list of consultants that saw the patient]
Course [summary of the patient's stay in the hospital]
Assessment_and_plan [summary of what's going on with the patient and the plan for next steps in the hospital or discharge]
Dc_instructions [includes follow-up appointments, diet and activity restrictions, labs to be done in the future]
Disposition [where the patient is going after discharge/transfer]
Code_status [whether or not the patient wants to be resuscitated]
Condition [usually one or two words describing the general state of the patient - e.g. "guarded" or "critical"]
Procedure_name [unique to radiology, pathology, and operative notes - says what procedure/operation was done]
Procedure_details [unique to radiology, pathology, echo, and operative notes - where they give lots of details about the operation or procedure including technical details, lots of measurements, etc.]
Complications [unique to operative notes - describe complications of the surgery that was performed]
Comments [unique to radiology, pathology, and echo reports - notes that the physician interpreting the study put in]

For retrieval, each section was mapped to the topic frame slots with various weights. Mostly, the best matching frame slot for a given document section was assigned a weight of 1.0, the less relevant slots were assigned a weight of 0.7 each, and the remaining slots were assigned the default weight of 0.1, with the exception of medications fields and allergies, for which weights are set to 0 for the mutually exclusive sections. A typical section to frame slots mapping is shown in table 2.

4. Experiments

Our experiments focused on finding ways to automate the use of domain knowledge that was shown to significantly improve retrieval results in the 2011 interactive runs. We established the baseline with an 'off-the-shelf' Lucene run (plain Lucene) and augmented Lucene and Essie with the same amounts of knowledge. As in 2011, we interactively modified Essie queries until the top 10 visits looked mostly relevant.

We used Lucene in several runs. We used plain Lucene for searching the positive and speculative text identified in the preprocessing of the visits. In addition, we used Lucene with two query generation approaches. In both approaches, we combined the original query with the query based on the topic frame ("generated query"). The original query was assigned a weight of 0.8 while the generated query got a weight of 1.0. The weights were estimated based on the TREC 2011 Medical Records track topics.

Table 2. Weights for ranking topic terms extracted into a frame slot in column 1 and found in the *History of present illness* and *Home medications* document section.

Topic Frame Slot	History of present illness	Home medications
Age	0.7	0.01
Gender	0.7	0.01
Population	0.7	0.01
PMH	0.7	0.01
SocialHx	0.01	0.01
AdmitProblem	0.7	0.01
DischargeProblem	0.01	0.01
Problem	0.7	0.01
Finding	0.7	0.01
ComplicationsOf	0.7	0.01
MedBeforeAdm	0.7	1
MedInHosp	0.7	0
MedOnDisch	0.01	0
Allergies	0.01	0
MedForPrblm	0.7	1
ProcForPrblm	0.7	0.01
Procedure	0.7	0.01
ProcFinding	0.7	0.01
FamilyHx	0.7	0.01
DischDest	0.01	0.01
CodeStatus	0.7	0.01
Location	0.7	0.01
ProcBeforeAdm	0.7	0.01

In the NLMLuceneExp run, we combined the original Lucene query and the topic frame representation of the query. The original query did not constrain the order of words but the generated query relied on phrase search for predicates. In specific cases, the query terms in a given expression were constrained to be found within a specific number of words using the character ~ followed by the maximum allowed length of the span of text. In addition, we performed expansion of terms based on the Google search strategy that we developed in 2011. The expansion was performed for the drugs and procedures entity types.

When the topic frames specified the ages of the patients, we searched the AGE fields generated during document preprocessing. Since the ages were not available for all visits, we could miss visits without the age field. So instead of looking for visits having the specific age range, we resorted to retrieving visits not having ages out of the expanded range. For instance, if we were looking for children, we excluded patients that were explicitly older than 12, but retained the patients for whom the age was not stated.

Overall we submitted the four runs described in Table 3.

Table 3 NLM runs submitted to the Medical Record Retrieval track

Run	Description
NLMManual	Interactively refined queries padded with an automatic run based on topic frames search over positive text in sections using Essie
EssieAuto	Essie search using topic frames over positive sections padded with lossy expansion
NLMLuceneSec	Uses the topic frames to identify relevant search terms and to do expansion (using Wikipedia and Google) of drug names and procedures. Search terms were weighted according to the section of the report in which they appear (as described in Table 2)
NLMLuceneExp	Uses topic frames to identify relevant search terms and to do expansion (using Wikipedia and Google) of drug names and procedures.

One experiment which proved not as helpful as we had hoped was to perform question expansion using NCBI's TexTool[1]. We sent the TexTool each of the 2011 topics and requested the top 200 PubMed Related Articles. We then summarized the MeSH Headings for the articles that TexTool found related to each topic by retaining only the MeSH Headings that occurred 100 or more times. These MeSH Headings were then manually reviewed to see if they would help to provide a broader view of the topic by expanding acronyms and by identifying potentially related terms. In the end, the existing query processing was richer and provided fewer ambiguity opportunities.

5. Results

Judging by the results of the Wilcoxon signed-rank test, our interactive Essie run was significantly better than all our other runs on all reported metrics.

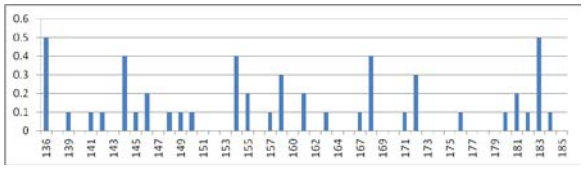

Figure 1. Differences in P@10 between the overall best and the NLM interactive run

Although we verified that most of the top ten documents in the NLMManual run were relevant, for several topics (that used few words to express the information needs) we were unsure that we had enough information to fully understand the relevance of a visit to the topic. Figure 1 shows the differences in P@10 between the overall best result for each topic and the NLM interactive run for the same topic. The large differences could partially be explained by the differences in interpretation of the information needs that are discussed in the next section.

The results of all our submitted automatic runs are in the same group. The NLMLuceneExp run appears to be significantly above the baseline for all metrics and significantly above the median for all metrics but P@10, whereas the remaining automatic runs are not above the baseline or the median. The results for all NLM runs are shown in Table 4.

Table 4 Evaluation results

Run	infAP	infDCG	R-prec	P@10
NLMManual	0.3663	0.68	0.477	0.7489
Lucene baseline	0.1664	0.4017	0.2909	0.4234
EssieAuto	0.1719	0.4042	0.282	0.4362
EssieAuto_bug_fix	0.1738	0.4154	0.2899	0.4617
NLMLuceneSec	0.1774	0.4414	0.3091	0.4745
NLMLuceneExp	0.1987	0.4649	0.3284	0.5043

After submitting the results, we found two bugs in EssieAuto topic processing. The queries for this run combined the topic frame-based queries with the original topics. The original topics contained parentheses that were not properly escaped and subsequently treated by the search engine as syntax errors, leading to low scores for the topics containing parentheses. Another bug in translation of the topic-frame age slot to queries caused low performance on queries that specified patients' ages. Fixing the bugs did not significantly improve the results for the EssieAuto run (EssieAuto_bug_fix in Table 4).

Looking at the NLMLuceneExp results, query expansion and proximity search seem to have a

[1] http://ii.nlm.nih.gov/MTI/related.shtml

positive effect on Lucene retrieval, while the NLMLuceneSec performance shows we still have to learn how to use the section information for a more effective retrieval.

6. Discussion

We will focus the discussion on the differences in interpretations of the topics that are manifested as the large differences for P@10 between the best score for the topic and our interactive run for this topic. Our interpretation was too strict for topic 136, *Children with dental caries*. We first interpreted the topic very strictly and established a hard constraint on the age field requiring its value to be "birth-12". This restriction resulted in very few relevant visits; we therefore expanded the age filter to include the "in teens" group as well. We did not consider patients who were clearly adults (based on the chart review) but without explicitly stated age to be more relevant than the patients that had their age stated. The relevance judgments however considered the adults without the explicitly stated age to be somewhat relevant.

Another example of our misinterpretation of the information needs is topic 154, *Patients with Primary Open Angle Glaucoma (POAG)*. Whereas we focused on POAG and its symptoms (increased intraocular pressure, eye pain, and blurry vision), anyone with a history of glaucoma was judged relevant to this topic. Table 5 shows our reasons to judge documents relevant to topic 154 and the actual relevance judgment.

Table 5. Reasons for judging documents relevant to topic 154

Visit ID	score	Reasons for our interactive relevance judgments
7tQ6HF6v7w9k	0	"Possible glaucoma" "RIGHT EYE PAIN AND INCREASED INTRAOCULAR PRESSURE."
TveRWQfhKhYx	0	The patient presents today with acute glaucoma after corneal transplant.
I7Dk9G/pCQbO	0	history of "elevated intraocular pressure bilaterally"
V0JfQ0OniN+P	1	eye pain, blurriness, incr pressure on exam

Finally, for topic 144, *Patients with diabetes mellitus who also have thrombocytosis*, the poor performance is due to a combination of differences in interpretation and our decision to use ICD-9 codes, as well as our inability to use numeric values. Using the UMLS synonymy, our search treated 'increased platelet count' and 'thrombocythemia' as synonyms of 'thrombocytosis', whereas as shown in Table 6, this is not the case for the relevance judgments. Table 6 shows the other disagreements on this topic's top documents retrieved by our interactive run.

Table 6. Reasons for judging documents relevant to topic 144

Visit ID	score	Reasons for our interactive relevance judgments
7cGF1R99z8jb	0	Has diabetes, portal vein thrombosis, essential thrombocythemia (synonym for thrombocytosis)
Ck/D5AD9G1GA	0	Has essential thrombocytosis, but diabetes is only in ICD9 codes, not in text
hI3wi7+RGLi/	0	Has diabetes, thrombocytopenia in text, thrombocytosis in ICD9
t4eF9N+9g3+V	0	Steroid induced diabetes, thrombocythemia ICD9

Looking at the judgments for topic 144 and topic 152, *Patients with Diabetes exhibiting good Hemoglobin A1c Control (<8.0%)*, we assume that being able to issue database-like queries for finding values above the normal platelet count range (150-450) and other queries requiring extraction and evaluation of numeric values might become a desirable feature for a search engine. None of our search engines or NLP tools is able to perform such ad-hoc queries, which opens an interesting direction for future research.

Overall, the analysis of our 2012 Medical Records track results is consistent with our 2011 observations: the search engines perform the traditional tasks as expected; the interactive query formulation significantly improves the results; a narrative description of information needs in addition to the tersely formulated topics could improve the interactive query formulation by explicitly stating the information needs and what will constitute potentially relevant documents. Finally, we hope the medical records track will continue and help us define the functions of the search engines and the points at which the search engines should hand the documents over to NLP tools or be combined with structured queries.

References

Demner-Fushman D, Abhyankar S, Jimeno-Yepes A, Loane R, Rance B, Lang FM, Ide N, Apostolova E, Aronson AR. Knowledge-based approach to medical records retrieval. The Twentieth Text Retrieval Conference TREC-2011, Gaithersburg, MD