Concept-centric Indexing and Retrieval on Medical Text

David Eichmann

Biomedical Informatics Group
Institute for Clinical and Translational Science
The University of Iowa

Introduction

The NIH Clinical and Translational Science Award (CTSA) program has resulted in the formation of new research interactions for many IR and NLP research groups. Research access to large-scale clinical data is proving to be a critical component of the overall goals of the CTSA. While much of the clinical record is tabular and structured, substantial amounts of pertinent information reside in unstructured text attached to those structured records. This is particularly true for research subject cohort identification, where the inclusion and exclusion criteria for a given study (e.g., family history, quality of life assessments, etc.) may not well align with the data captured in a typical clinical encounter. The TREC Medical Record track provides an excellent means to drive innovation in clinical data retrieval, particularly for unstructured elements of the electronic medical record.

Approach

Our ongoing interactions with clinical researchers seeking access to our data warehouse strongly confirms the conceptual, categorical nature of the sample queries provided to track participants – 'atypical antipsychotics' is a category of medication and not a specific medication and medical records almost exclusively list specific medications (at specific doses). This resulted in our core hypothesis for our architecture – perform concept recognition and extraction from both documents and queries with hierarchical downward expansion of query concepts to match against the specifics of concepts mentioned in documents. This resulted in the following phases of processing for the corpus:

- XML parsing and segmentation. Analysis of the collection indicated recurring high levels of structure for elements such as medications and problem lists. We handle these separately from free text.
- Part-of-speech tagging and sentence boundary detection for those document segments appearing to be unstructured text.
- UMLS concept extracted at the sentence level, using bi-directional greedy dictionary matching for noun phrases.
- Negation recognition at the sentence level. We used a variant of the NegEx algorithm [1] to flag sentences as likely carrying negated concepts.

Our experience with clinical researchers also led us to attempt to identify gender, ethnicity and age (actually decade of age, given the data) of the patient, typically based upon the opening sentence of the report. Tables 1a-c reflect the results of this extraction. While these attributes would normally be available as attributes of the structured elements of the EMR, we assumed it necessary to detect these when possible as they were potential inclusion/exclusion criteria for a topic.

Gender	Count
female	12588
male	14260
<null>	74018

Table 1a. Detected gender frequency.

Ethnicity	Count
black	174
white	4806
<null>	95886

Table 1b. Detected ethnicity frequency.

Age	Count
5	555
10	1218
20	3897
30	3673
40	5622
50	7130
60	7503
70	7555
80	7335
90	1400
<null>	54978

Table 1c. Detected age frequency.

We then processed the queries in a similar manner, yielding a set of concepts for each query. Each concept was then expanded by inclusion of any concepts appearing below the concept of interest in the UMLS hierarchy, capped at a maximum of 100 expansion concepts per original concept. Retrieval was then done using disjunctive matching of all concepts in the query against the aggregated set of all concepts for a visit (i.e., all sentences for all documents for the given visit). Scoring was done either using the total number of matches or the count of distinct concept matches. Our submitted runs were hence a 2 • 2 • 2 cube of the following configuration parameters:

- Scoring by total number of concept matches or distinct concept matches
- Use of ICD-9 diagnosis codes from the document headers or not
- Use of negation flags to suppress concept matches or not

Our first round of submissions for pooling involved the four parameter permutations with no negation suppression. The second round of submissions involved the four parameter permutations with negation suppression. Run configurations are shown in Table 2 below.

Run	Sum Scoring	ICD9 Used	Negation Excluded	Judged
UIICTSmed01	F	F	F	T
UIICTSmed02	T	F	F	T
UIICTSmed03	F	T	F	T
UIICTSmed04	T	T	F	T
UIICTSmed05	F	F	T	F
UIICTSmed06	T	F	T	F
UIICTSmed07	F	T	T	F
UIICTSmed08	T	T	T	F

Table 2. Parameters used in submitted runs.

Results

As noted by the track organizers, R-prec, bpref and P@10 (all precision measures) were used due to difficulty in running the evaluation program against submissions using the low numbers of judgments for topics. R-Precision measures precision after R docs have been retrieved, where R is the total number of relevant docs for a query. [2] bpref uses binary relevance judgments to define the preference relation (any relevant document is preferred over any nonrelevant document for a given topic). [3] We include *est_recall* – a simple, and admittedly ad hoc, means of judging recall by dividing the number of relevant documents for a topic into the number relevant returned by the system. While clearly not properly a measure of recall against the corpus, it at least provides us a means of comparison to other system in the evaluation by including all (judged) relevant documents found by all systems. Hence we show in Table 3 our eight submitted runs and their results.

Run	num_ret	num_rel	num_rel_ret	R-prec	bpref	P@10	*est_recall*
UIICTSmed01	29031	1765	1405	0.2285	0.3700	0.3618	0.7960
UIICTSmed02	29031	1765	1129	0.1541	0.2822	0.2500	0.6396
UIICTSmed03	29161	1765	1422	0.2372	0.3935	0.3412	0.8056
UIICTSmed04	29161	1765	1156	0.1630	0.3090	0.2059	0.6549
UIICTSmed05	28863	1765	1405	0.2301	0.3727	0.3441	0.7960
UIICTSmed06	28863	1765	1132	0.1579	0.2847	0.2529	0.6413
UIICTSmed07	28999	1765	1421	0.2380	0.3954	0.3206	0.8050
UIICTSmed08	28999	1765	1155	0.1618	0.3103	0.2118	0.6543

Table 3. Summary results for all submitted runs.
Note that est_recall is an estimated recall derived as num_rel_ret / num_rel.

Table 4 shows the R-prec performance of all runs, in descending R-prec order.

Run	R-prec	Sum Scoring	ICD9 Used	Negation Excluded
UIICTSmed07	0.2380	F	T	T
UIICTSmed03	0.2372	F	T	F
UIICTSmed05	0.2301	F	F	T
UIICTSmed01	0.2285	F	F	F
UIICTSmed04	0.1630	T	T	F
UIICTSmed08	0.1618	T	T	T
UIICTSmed06	0.1579	T	F	T
UIICTSmed02	0.1541	T	F	F

Table 4. Runs and parameters ordered by R-prec.

Simiarly, Table 5 shows the bpref performance of all runs, in descending bpref order.

Run	bpref	Sum Scoring	ICD9 Used	Negation Excluded
UIICTSmed07	0.3954	F	T	T
UIICTSmed03	0.3935	F	T	F
UIICTSmed05	0.3727	F	F	T
UIICTSmed01	0.3700	F	F	F
UIICTSmed08	0.3103	T	T	T
UIICTSmed04	0.3090	T	T	F
UIICTSmed06	0.2847	T	F	T
UIICTSmed02	0.2822	T	F	F

Table 5. Runs and parameters ordered by bpref.

Table 6 shows the P@10 performance of all runs, in descending P@10 order.

Run	P@10	Sum Scoring	ICD9 Used	Negation Excluded
UIICTSmed01	0.3618	F	F	F
UIICTSmed05	0.3441	F	F	T
UIICTSmed03	0.3412	F	T	F
UIICTSmed07	0.3206	F	T	T
UIICTSmed06	0.2529	T	F	T
UIICTSmed02	0.2500	T	F	F
UIICTSmed08	0.2118	T	T	T
UIICTSmed04	0.2059	T	T	F

Table 6. Runs and parameters order by P@10.

Finally, Table 7 shows the estimated recall performance of all runs, in descending est_recall order.

Run	*Est. Recall*	Sum Scoring	ICD9 Used	Negation Excluded
UIICTSmed03	0.8056	F	T	F
UIICTSmed07	0.8050	F	T	T
UIICTSmed01	0.7960	F	F	F
UIICTSmed05	0.7960	F	F	T
UIICTSmed04	0.6549	T	T	F
UIICTSmed08	0.6543	T	T	T
UIICTSmed06	0.6413	T	F	T
UIICTSmed02	0.6396	T	F	F

Table 7. Runs and parameters ordered by estimated recall.

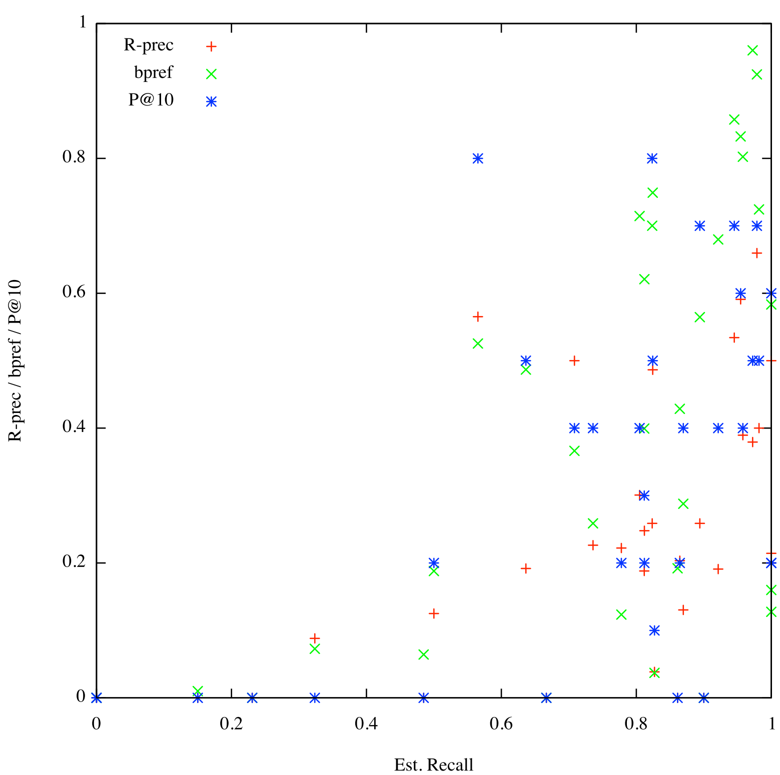

Figure 1. Estimated recall vs. evaluation measures by topic

Topic	Num Ret	Num Rel	Num Rel Ret	R-prec	bpref	P@10	Est. Recall
101	432	74	61	0.4865	0.7491	0.5000	0.8243
102	1000	89	82	0.1910	0.6798	0.4000	0.9213
103	1000	12	12	0.5000	0.5833	0.6000	1.0000
104	1000	9	7	0.2222	0.1235	0.2000	0.7778
105	958	145	141	0.3793	0.9602	0.5000	0.9724
106	1000	85	69	0.1882	0.3994	0.3000	0.8117
107	17	23	13	0.5652	0.5255	0.8000	0.5652
108	1000	13	3	0.0000	0.0000	0.0000	0.2307
109	1000	123	99	0.3008	0.7145	0.4000	0.8048
110	1000	95	91	0.3895	0.8024	0.4000	0.9578
111	1	21	0	0.0000	0.0000	0.0000	0.0000
112	1000	73	69	0.5342	0.8576	0.7000	0.9452
113	1000	14	14	0.2143	0.1276	0.2000	1.0000
114	1000	55	54	0.4000	0.7243	0.5000	0.9818
115	1000	36	31	0.0000	0.1921	0.0000	0.8611
116	1000	10	9	0.0000	0.0000	0.0000	0.9000
117	41	22	21	0.5909	0.8326	0.6000	0.9545
118	1000	52	43	0.0385	0.0370	0.1000	0.8269
119	1000	46	40	0.1304	0.2878	0.4000	0.8695
120	1000	117	95	0.2479	0.6209	0.2000	0.8119
121	1000	40	20	0.1250	0.1881	0.2000	0.5000
122	1000	24	17	0.5000	0.3663	0.4000	0.7083
123	1000	33	16	0.0000	0.0643	0.0000	0.4848
124	1000	6	4	0.0000	0.0000	0.0000	0.6666
125	271	14	0	0.0000	0.0000	0.0000	0.0000
126	279	5	5	0.2000	0.1600	0.2000	1.0000
127	1000	85	70	0.2588	0.7001	0.8000	0.8235
128	1000	85	76	0.2588	0.5646	0.7000	0.8941
129	1000	53	39	0.2264	0.2588	0.4000	0.7358
131	1000	99	63	0.1919	0.4868	0.5000	0.6363
132	1000	94	92	0.6596	0.9245	0.7000	0.9787
133	1000	20	3	0.0000	0.0100	0.0000	0.1500
134	1000	34	11	0.0882	0.0727	0.0000	0.3235
135	1000	59	51	0.2034	0.4286	0.2000	0.8644
all	28999	1765	1421	0.2380	0.3954	0.3206	0.8050

Table 8. By-topic performance for UIICTSmed07.

Discussion

Table 8 shows by by-topic performance for UIICTSmed07, representative of the set of runs as a whole. Figure 1 plots est_recall against the three evaluation measures. While our system was tuned for recall to match the semantics of the task, performance against the precision-focused measures is frequently quite respectable. The R-prec values will be particularly interesting to explore. We made no particular attempt to rank visits by anything other than a quite coarse metric (number of matched UMLS concepts). Given the est_recall values, it might be possible to substantially enhance R-prec through modest attention to the ranking algorithm.

The patterns of performance for three parameters show interesting similarity between R-prec, bpref, and est_recall. With one minor exception in order, the various system configurations function relatively the same across the three measures. Using the count of distinct occurrences of a concept clearly outperforms total occurrences. Interestingly, the ICD-9 codes in the report metadata provide no significant benefit in scoring. This is likely due to those metadata being generated by human coders from the text of those same reports. Excluding concept-mentioning sentences involving negation also appears to have had negligible impact, positively or negatively. We intend to explore negation more fully in our future analyses.

References

1. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*. 2001;34:301-10
2. http://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/trec_eval_video/README.
3. Buckley, C and Voorhees, E. Retrieval Evaluation with Incomplete Information. *SIGIR'04*, July 25–29, 2004, Sheffield, South Yorkshire, UK.