Empirical Ontologies for Cohort Identification

Stephen Wu, Kavishwar Wagholikar, Sunghwan Sohn, Vinod Kaggal, Hongfang Liu
Mayo Clinic, Rochester, MN

1 Introduction

The growth of patient data stored in Electronic Medical Records (EMR) has greatly expanded the potential for the evidence-based improvement of clinical practice. The proper re-use of this clinical information, however, does not replace basic research techniques — it augments them. The Text REtrieval Conference 2011 Medical Records Track explored how information retrieval may support clinical research by providing an efficient means to identify cohorts for clinical studies.

Mayo Clinic NLP's submission to the TREC Medical Records track attempts information retrieval at a semantic level, combining two disparate means of computing clinical semantics. Substantial effort has gone into the development of precise semantic specification of concepts in medical *ontologies* and terminologies[1, **?**]. But human clinicians do not generate clinical text by referring to such resources, and ontology creators do not base their terminology design on clinical text — so the *distribution* of ontology concepts in actual clinical texts may differ greatly.

Therefore, in representing clinical reports for cohort identification, we advocate for a model that makes use of expert knowledge, is empirically validated, and considers context. This is accomplished through a new framework: *empirical ontologies*. Patient cohort identification is thus a practical use case for the techniques in our recent work on clinical concept frequency comparisons[2, 3].

The rest of this paper describes the TREC 2011 Medical Records task, describes Mayo Clinic's run submissions, and reports evaluation results with subsequent discussion.

2 Background

The inaugural TREC 2011 Medical Records track was arranged as follows. The data to be retrieved lay in the University of Pittsburgh's BLU repository, which includes only the free text portions of medical records. Each patient at the University of Pittsburgh would have one or more medical *records* (documents) associated with him or herself. Each record was given in XML format, and included both structured data and the unstructured text.

```
<?xml version=''1.0'' encoding=''UTF-8''
standalone=''no''?>
<report>
<checksum>20060201ER-Fs2xiJYPXwVE-848-1341620775
</checksum>
<subtype>EVAL</subtype>
<type>ER</type>
<chief_complaint>DENTAL
PAIN</chief_complaint>
<admit_diagnosis>521.00</admit_diagnosis>
<discharge_diagnosis>525.9,E917.9,
</discharge_diagnosis>
<year>2007</year>
<downlaod_time>2008-02-06</downlaod_time>
<update_time/>
<deid>v.6.22.06.0</deid>
<report_text>[Report de-identified
(Safe-harbor compliant) by De-ID
v.6.22.06.0]
.
.
.
</report_text> </report>
```

Records are uniquely identified by their `checksum`. Note that each record contains a note `type` and `subtype`; in the example, the note comes from an Emergency Room/Department. The `chief_complaint` section is a helpful textual summary of what the record is about from the patient's perspective, and is not present for every record. The `admit_diagnosis` and `discharge_diagnosis` serve a similar function but are also not always present. They are given as ICD-9 codes, a medical terminology frequently used for billing purposes. Finally, notice that the notes were de-identified, so that any protected health information has been replaced with surrogates.

The records were grouped into *visits* — a physical visit to the hospital. The unit of retrieval was defined as a patient visit. In total, there were 95,702 records

that corresponded to 17,198 visits. The largest visit was 418 records, but the mean visit was 5.56 records.

Participants from 29 institutions were given a set of 35 hypothetical topics developed by experts at the Oregon Health Sciences University (OHSU). These topics defined patient profiles that might be involved in a clinical trial. For each topic, participants retrieved a list of patient visits in order of relevance to the topic.

For evaluation and ranking, there were two evaluation rounds. In the *judged* round, retrieved records from participants' runs were given to assessors at OHSU. These assessors rendered relevance judgments on a stratified pool of visits — the top 10 of each submitted run, and decreasing percentages of lower-ranked visits from each run.

Although there was a fairly consistent pool depth (number of adjudged visits for each topic), the nature of each topic and its correspondence with the given dataset varied greatly. For example, topic 130 was discarded for purposes of evaluation because no records were assessed as being relevant to the query topic; on the other hand, other topics likely had many relevant visits that were never assessed.

Thus, the primary evaluation metric chosen for rankings was bpref, since it only penalizes systems for placing irrelevant documents ahead of relevant documents. Other evaluation metrics such as P@10 will be presented where they illustrate something about the data, system outputs, or evaluation procedure.

In the *unjudged* round of evaluations, bpref and other metrics were calculated based on the relevance judgment pools from the first round. Thus, a metric like P@10 for an unjudged round is an approximation, unlike that for a judged round.

Mayo Clinic NLP submitted runs to both the judged and unjudged rounds, but we report only the latter (the former contained significant bugs).

3 Methods

In our empirical ontologies approach, each query is predominantly represented at the semantic concept level, rather than at a textual level. Similarly, each document is represented as a set of semantic concepts. Query concepts and document concepts are weighted, aggregated, and compared to produce a ranking of the most similar topic–visit pairs.

Below, we describe Mayo Clinic's baseline unjudged run and then detail the innovations introduced in the other runs.

3.1 Baseline

Each topic query is given in a form such as

> Number: 108
> Patients treated for vascular claudication surgically

From the text of this query, we manually assigned semantic concepts to important semantic concepts. As a baseline configuration (outside the dotted line of Figure 1), the weights for each query were specified manually. Three medical experts were asked the following questions:

1. What are essential concepts in the query?
2. What procedures, medication, or surgeries treat this (if not in the topic description)?
3. What distinction is being drawn (things that are NOT wanted)?
4. What is a typical patient with this condition like? Common comorbidities?
5. What types of notes will this be found in? NOT found in? (e.g., discharge diagnosis, operative note)
6. Are there other terms that express what they are looking for better?

The answers to these questions were at first manually written as term lists by the experts. We normalized these terms to concept unique identifiers (CUIs) from the the Unified Medical Language System (UMLS) Metathesaurus[1] via the UMLS Terminology Services. We augmented these CUIs with pseudo-CUIs for age and gender terms, as well as terms from the last question that might be in text but might not be capturable by dictionary lookup algorithms.

Weights for each of the concepts were given a weight according to which question they were answering: 1.0, 0.75, 0.5, and 0.25 for the first 4 questions, respectively. CUIs for item 3 were marked as negated. The second-to-last question was unused for the baseline method, but, for use in later methods, weights were given to note types for each topic. Beyond the questions we asked experts, weights were also given to sections within a medical record upon analysis of the topics; these were also unused in the baseline method.

We then view each topic query as a vector, where each index indicates the weight of a concepts. For the baseline model, the query processing ends here, and we have produced a query 'mask.'

Figure 2 shows the methodology for processing and retrieving reports, where again, the baseline method

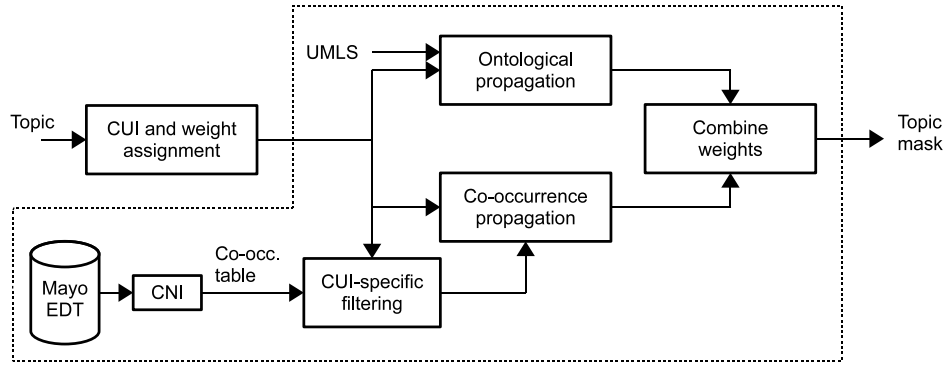

Figure 1: Block diagram for query expansion. Dotted line separates baseline and other techniques.

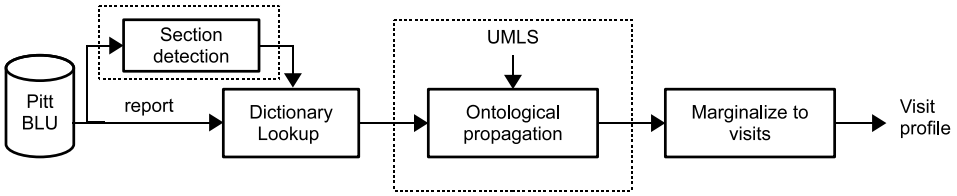

Figure 2: Block diagram for visit summarization. Dotted line separates baseline and other techniques.

is outside the dotted line. First, we automatically extract concepts from reports in the Pittsburgh repository using standard clinical NLP tools. In the baseline method, this mapping from text to UMLS concepts was accomplished with an Aho-Corasick[4] dictionary lookup on normalized tokens, using the UMLS 2011AA as the dictionary.

For report-side processing, we again had a few special cases, where we automatically assigned CUIs according to rules. ICD-9 diagnosis codes were translated into CUIs via the Metathesaurus, and were included. Special pseudo-CUIs were created for age and gender groups. Also, the same pseudo-CUIs from the query side (age, gender, terms from the last question) were included.

For each report, we count the frequency of all concepts (CUIs) that are present. Our previous work shows that this is relatively similar to tf-idf in the medical domain[3]. Then, for each hospital visit, weights are summed across reports to obtain a vectorial representation. This allows us to represent the patient's visit in vectorial form, just as we had represented the query.

We can then rank query–visit pairs by comparing the vectors. The simplest method for doing this, cosine similarity, was used as our baseline. The 1,000 visits closest to each query topic vector were ranked and retained as our run submission.

3.2 Propagation

The idea of empirical ontologies, while present in the baseline system, is implemented most fully with the idea of propagation. Because we are viewing each query and each report as a collection of weighted semantic concepts, we can make additional inferences at a semantic level. These are illustrated in Figure 3.

Ontological propagation. The UMLS Metathesaurus contains CUIs that arise from source ontologies, which maintain hierarchical relationships between concepts. For example, an "colon structure" (CUI: C0009368) has an 'isa' relationship with its ontological parent, "large intestine structure" (CUI: C1709915). Ontological propagation takes a set of weighted concepts (i.e., directly from the queries, or from frequency counts in the reports) and propagates those weights to hierarchically related concepts.

If a concept "colon structure" is mentioned, the presence of its parent "large intestine structure" is entailed. Thus, we propagate weights from a concept to its parents (see Figure 3 for the corresponding example). However, if a concept "colonoscopy" is negated "no colonoscopy," it does not imply that no procedures were done. However, it may imply that "no virtual colonoscopy" was done. Therefore, for negated concepts, we propagate weights from a concept to its children.

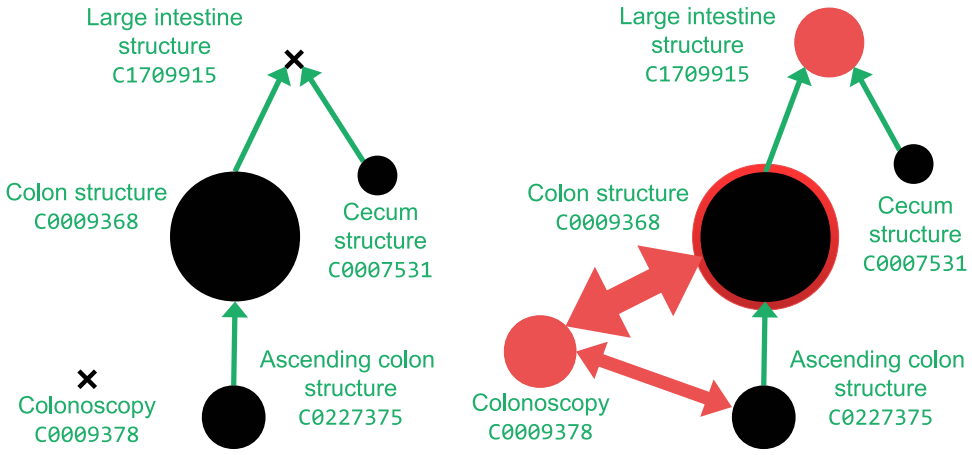

Figure 3: Left: Hierarchically-arranged concepts with weights represented by circle sizes. 'x' denotes a missing concept. Right: Propagation of a concept ("colon structure") to hierarchically related concepts ("large intestine structure") and co-occurrence related concepts ("colonoscopy"). The red (shaded) circles indicate weights added by the propagation. The strength of co-occurrence relationships are shown as the size of red arrows.

Of course, it is not always meaningful to execute this kind of propagation fully. For example, it is may not be helpful to say that there is an instance of a "anatomical site" when the concept "colon structure" is discovered. Therefore, we implement two kinds of decay and allow an optional cutoff. First, we included a constant multiplier (less than 1) to make any propagated concept count less than an originally-discovered concept. Second, we implement a geometric decay based on path distance (number of hierarchical nodes separating the original weighted concept and the concept that will receive additional weighting). Finally, we give the option to cut off upwards or downwards propagation at defined path distances.

Figures 1 and 2 show where Ontological propagation takes place in our system.

Co-occurrence propagation. Ontological semantic relationships, as coded by experts, sometimes ignore the empirical realities of how concepts are distributed. We also investigated distributional relationships between concepts, so that concepts occurring in the same document would have relationships between them with some weight defined by how many times they co-occurred. This is shown on the right of Figure 3 by the size of the arrows between "colon structure" and "colonoscopy" and between "colon structure" and "ascending colon structure."

We created two large tables storing the frequency of CUI occurrences in each document, based on a random sample of 50 million Mayo Clinic clinical notes. The smaller had 79,597,219 relationships, and the larger had 282,712,288. These tables excluded the most common concepts in the clinical documents due to computational constraints. The intuition was that common concepts are unlikely to be highly discriminative, since previous work shows they are fairly general terms[3].

These tables were queried in SQL to retrieve co-occurrences, as follows:

```
SELECT a.cui,b.cui,COUNT(*)
FROM doccuifreq a, doccuifreq
b WHERE a.cui=$incui AND
a.doc=b.doc AND a.cui<>b.cui
GROUP BY b.cui HAVING
count(*)>cutoff
```

where `cutoff` was the fewest co-occurrences that would be reported as a valid relationship. Because searching for the co-occurrences of a concept is time-consuming, we limited this search to only the query expansion section, i.e., to concepts that were present in the query.

3.3 Variants

Mayo Clinic NLP submitted 4 system configurations:

1. MAYO2NOPROP (Section 3.1): Baseline configuration with manually-assigned query concepts and weights; report weights are based on Aho-Corasick string matching.

2. MAYOLBRST (Section 3.3.1): Includes the additional processing from the dotted boxes of Figures 1 and 2.

3. MAYOLBRA (Section 3.3.2): Replaces MAYO2NOPROP's maual query weights with automatic weights from Aho-Corasick string matching on the text of the query.

4. MAYOUBR (Section 3.3.3): Uses cTAKES instead of Aho-Corasick for Dictionary Lookup on reports.

3.3.1 Expansion and weighting (MAYOLBRST)

The configuration of the MAYOLBRST run is as follows:

- Query-side ontological propagation. Weights of query concepts are extended to UMLS 'isa' relationships (ontological neighbors). The initial downweighting of ontologically propagated concepts was set at 0.9 for positive concepts and 0.7 for negative concepts. The geometric decay parameter based on the path was set at 0.125.

- Query-side co-occurrence propagation. Weights of query concepts are extended to co-occurring concepts from Mayo's NLP-processed corpus of 50 million-plus clinical notes.

- Report-side ontological propagation. Weights of report concepts are extended to UMLS 'isa' relationships (ontological neighbors). The same decay parameters were used as the query-side ontological propagation above, but additionally cut off propagation towards parents at a path distance of 3, and towards children at a path distance of 2.

- Type weights. A multiplier is included on the report side for what type of note each concept is from, since a visit may have reports of different types. For example, for Topic 108 (see Section 3.1), concepts found in Operative notes were deemed 1.5× more useful than those found in other note types.

- Section weights. A multiplier is included on the report side for what section of a note each concept is from, e.g., DIAGNOSIS might be higher weighted than FAMILY HISTORY. Sections are automatically tagged using SecTagger, as per Figure 2.

3.3.2 Automatic query-side concept extraction (MAYOLBRA)

This run, MAYOLBRA, differed from the other 3 in the means by which it interpreted the query topic. The others used manual CUI assignments and weights, whereas MAYOLBRA assigned these CUIs and weights automatically. The same baseline dictionary lookup procedure (based on the Aho-Corasick algorithm) that was used on the reports was also used on the raw text of the query.

This implied that query processing had no special weighting, since the concept extraction had methodology used frequency, and the concept frequency in the queries was not necessarily tied to importance. However, this also meant that the same CUI(s) would be found from both the queries and the reports for a given text string.

The propagation configuration of this run was identical to that of MAYOLBRST. No type weights or section weights were used.

3.3.3 UIMA-based report-side concept extraction (MAYOUBR)

This run, MAYOUBR, differed from the other 3 in the means by which concepts were found in the reports. Report processing was carried out using Mayo Clinic's clinical Text Analysis and Knowledge Extraction System, cTAKES[5]. The query-side processing was manual, consistent with the baseline. This alternative report-processing pipeline was used out-of-the-box, and no tuning was added for age, gender, special terms, or ICD-9 diagnostic codes.

The propagation configuration of this run was again identical to that of MAYOLBRST. No type weights or section weights were used.

3.4 Development Evaluation

During development, we evaluated which of our algorithms might be useful by performing case-insensitive string matching of terms from each query, and then examining the top N reports. We then counted the percentage of records that contained matches. This metric was averaged across queries for each of our submitted runs, yielding a % accuracy measure similar to P@N that tests our methods against unweighted string matching. It is a rough measure, since there is no test to ensure that all the conditions of the query are met. The accuracy for the top 10, 20, 50, and 100 reports are:

```
1. mayo2noprop: 90.29, 88.86, 87.26, 85.51
2. mayolbrst:   90.57, 89.57, 87.26, 84.83
```

```
3. mayolbra:    68.85, 68.57, 66.40, 64.65
4. mayoubr:     80.57, 79.42, 76.97, 75.00
```

The 4 runs were chosen based on the development evaluations. Preliminary results showed that expanding queries and reports based on ontologies and co-occurrences did not make a significant difference in the accuracy. This is not unexpected; extending the weights to related concepts would not necessarily increase the number of exact string matches (which were used to define the metric). The official results showed that this was not the case on a real relevance judgment task.

Other variations were tested, but unsubmitted for final runs. Using the developmental evaluation, we made a few additional observations. First, we varied the size of co-occurrence tables used. Recall that we stored to tables from which we calculated co-occurrence relationships; the smaller co-occurrence corpus surprisingly did not impact accuracy greatly. Additionally, we manually tested a few parameter settings for coefficient, decay, and cutoff in the propagation steps. The final configuration submitted in MAYOLBRST, MAYOLBRA, and MAYOUBR performed the best on the development evaluation task.

4 Evaluation

Official TREC results on the baseline and variants are as follows:

	bpref	R-prec	P@10
mayo2noprop	0.3930	0.1709	0.2353
mayolbrst	0.4260	0.2203	0.2794
mayolbra	0.4527	0.2222	0.2794
mayoubr	0.2503	0.1059	0.1824

Ranked according to bpref, this places MAYOLBRA in 13th place out of 29 participating institutions at TREC 2011.

5 Discussion

It is evident that there was a significant effect of including the more sophisticated query expansion features. Interestingly, the automatic system receieved the best score on all metrics. This is especially interesting because our accuracy metric used for development did found the opposite result. This is likely due to the fact that when the query was processed with the same dictionary lookup algorithm as the reports, and the exact same CUIs were used for both, but this did not necessarily correspond to exact string matches in the development metric.

It is also clear from comparing MAYOUBR to MAYOLBRST that what goes into the report processing makes a significant difference, though the fact that metadata was unused in mayoubr makes the result inconclusive for the UIMA-based concept extraction approaches in general. Future versions with cTAKES could increase in accuracy considerably.

It should be noted that these were in the "unjudged" competition, and a large proportion of the returned reports were therefore unjudged. Some of the metrics (especially P@10) are dominated by the fact that these were unjudged runs. In each of the 4 systems, less than 50% of the top 10 visits were judged. Thus, the reported values of P@10 were close to the lowest possible values (if all the unjudged visits turned out to be irrelevant). This metric is particularly susceptible to the lack of judged visits; indeed, a meta-analysis of all submitted TREC runs showed that the P@10 for our systems were abnormally low compared to others with similar bpref scores.

6 Conclusion

The TREC 2011 Medical Records track competition provided an opportunity to use the semantically-oriented empirical ontologies for a practical information retrieval task, cohort identification. We obtained competitive results that illustrated the usefulness of advanced features like propagating weights between related concepts. Additionally, we found that the match between query-side and report-side algorithms had the most significant effect on performance.

Future work includes viewing the information retrieval task from a text-centric perspective without losing some of the gains that are possible from semantic reasoning.

References

[1] D.A. Lindberg, B.L. Humphreys, and A.T. McCray. The unified medical language system. *Methods of information in Medicine*, 32(4):281, 1993.

[2] Stephen Wu and Hongfang Liu. Semantic Characteristics of NLP-extracted Concepts in Clinical Notes vs. Biomedical Literature. In *Proceedings of AMIA 2011*, 2011.

[3] Stephen Wu, Hongfang Liu, Dingcheng Li, Cui Tao, Mark Musen, Christopher Chute, and Nigam Shah. UMLS Term Occurrences in Clinical Notes: A Large-scale Corpus Analysis. In *Proceedings*

of the AMIA Joint Summit on Clinical Research Informatics, 2012.

[4] Alfred V. Aho and Margaret J. Corasick. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6):333–340, 1975.

[5] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, and C.G. Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507, 2010.