Methodology for Creating UMLS Content Views Appropriate for Biomedical Natural Language Processing

Alan R. Aronson, PhD, James G. Mork, MSc, Aurélie Névéol, PhD, Sonya E. Shooshan, MLS, and Dina Demner-Fushman, MD, PhD

Lister Hill National Center for Biomedical Communications (LHNCBC)
U.S. National Library of Medicine, Bethesda, MD 20894

Abstract

Given the growth in UMLS Metathesaurus content and the consequent growth in language complexity, it is not surprising that NLP applications that depend on the UMLS are experiencing increased difficulty in maintaining adequate levels of performance. This phenomenon underscores the need for UMLS content views which can support NLP processing of both the biomedical literature and clinical text. We report on experiments designed to provide guidance as to whether to adopt a conservative vs. an aggressive approach to the construction of UMLS content views. We tested three conservative views and two new aggressive views against two NLP applications and found that the conservative views consistently performed better for the literature application, but the most aggressive view performed best for the clinical application.

INTRODUCTION

The Unified Medical Language System® (UMLS®) Knowledge Sources [1] contain a wealth of information that has been used to support biomedical Natural Language Processing (NLP) applications for many years. However, as the UMLS (and in particular the Metathesaurus®) has grown, the task of effectively using the Metathesaurus knowledge has grown more challenging.

Several Lister Hill NLP programs gain access to the knowledge embedded in the UMLS via the MetaMap program [2]. MetaMap employs three data models that differ in how much Metathesaurus content is filtered out [3]. The relaxed model filters out lexically similar strings based on case and hyphen variation, possessives, comma uninversion, *NOS* variation and non-essential parentheticals. It also includes the manual removal of some strings such as numbers, single alphabetics, *NEC* terms, Enzyme Commission (EC) terms, the short forms of brand names and, most importantly, unnecessarily ambiguous terms [4]. MetaMap's moderate model additionally filters out terms with certain term types, many of an abbreviatory nature. Finally, the strict model also filters out strings with complex structure; these are strings which MetaMap does not map well anyway. Over 40% of Metathesaurus strings are removed in the creation of the strict model. It is MetaMap's default model for semantic NLP processing, and it has been available as the first *Content View* since the 2005AA UMLS release [5].

Although the MetaMap strict data model supports NLP processing moderately well, it is far from perfect, especially with regard to ambiguity. And the problem has gotten progressively worse as more vocabularies have been added to the Metathesaurus. With this in mind, several Lister Hill researchers formed the Lister Hill NLP Content View (LNCV) project with the goal of constructing maintainable NLP content views of the Metathesaurus consisting of a biomedical literature view and multiple clinical views. This paper focuses mainly on the literature view.

Two major approaches emerged from early discussions in LNCV project meetings. The *conservative* approach, of which MetaMap data models are an example, consists of progressively removing Metathesaurus strings that are determined to be inappropriate for the view being constructed. The *aggressive* approach consists of wholesale removal of possibly detrimental strings followed by the restoration of strings that are determined to be appropriate, i.e., a *backoff* phase.

METHODS

In order to provide guidance as to which of the approaches, conservative or aggressive, to take in constructing a UMLS content view for the biomedical literature, we devised an experiment, reported here, in which we tested three conservative

Metathesaurus data views and two aggressive views against two NLP applications:

- NLM's Medical Text Indexer (MTI) [6,7]; and
- A clinical Problem and Intervention extraction program [8].

Metathesaurus data views

Three conservative Metathesaurus data views were extracted from the 2007AB Metathesaurus (English strings only) using progressively more MetaMap-style filtering: Base, AutoFilter and AllFilter. The Base view is simply MetaMap's relaxed data model. Because lexical filtering such as case variation and comma uninversion are so ingrained in MetaMap's behavior, this is as close as we could get to an "out-of-the-box" Metathesaurus data view. The AutoFilter view consists of all of MetaMap's automatic filtering; it is MetaMap's strict model but without the manual ambiguity filtering. And the AllFilter view is MetaMap's strict model. We included both AutoFilter and AllFilter to assess the value of the laborious ambiguity filtering we perform annually.

Two aggressive data views, Aggressive and AggrBackoff, were constructed based on substring matching respecting word boundaries but ignoring case variation. The Aggressive view, the most restrictive of all the views, consists of removing all 2007AB Metathesaurus strings that are a proper substring of another string in the same concept, respecting word boundaries. The aggressive views arose from the observation that some substrings are inappropriate representatives for their concept. For example, *malaria* occurs in the concept *Malaria Vaccines*, and *resistance* occurs in *social resistance*. Conversely, an example in which the removal of a substring is not useful is that *Alzheimer* is removed from the concept *Alzheimer's Disease*.

As a way of overcoming the overaggressiveness of the Aggressive view, the AggrBackoff view was constructed from it by restoring strings *X* having a superstring in its concept containing *X* and one of five words (or their plurals) to its right: *disease(s)*, *syndrome(s)*, *disorder(s)*, *protein(s)* or *gene(s)*. Again, word boundaries are respected and case is ignored. In the Alzheimer's Disease case above, *Alzheimer* is restored because of either *Alzheimer syndrome* or *Alzheimer's Disease*.

The order of the Metathesaurus views from most conservative to most aggressive is Base, AutoFilter, AllFilter, AggrBackoff and Aggressive.

LNCV document collection

The set of documents used in our experiments, the LNCV document collection, consists of a randomly chosen subset of size 10,000 from the set of approximately 650,000 MEDLINE citations that were indexed in 2007 and had MTI recommendations. Completion dates for the collection range from mid-November 2006 to mid-November 2007. The composition of the LNCV collection with regard to presence or absence of an abstract is about the same as for the entire collection with 83% of the citations having abstracts.

MTI experiment

The MTI experiment consisted of processing the LNCV document collection through MTI [6,7] replacing the normal MetaMap data model with one of the five Metathesaurus data views defined above. The indexing recommendations so obtained were compared with the official MeSH indexing for the documents, computing Recall (R), Precision (P) and F_2 values for each document. The F-measure $F_2 = 3*(RP)/(R+2P)$ gives Recall twice as much weight as Precision in order to reflect the indexing perspective that finding additional relevant indexing terms is more important than including a few irrelevant terms.

Problem and Intervention extraction experiment

A second experiment involved processing the LNCV document collection against a problem and intervention extraction facility. The problem and intervention extractors identify two of the four elements of a well-built clinical query of MEDLINE abstracts, which is used to organize knowledge structure in the Clinical Question Answering (CQA) decision support system [8].

The problem extractor relies on recognition of concepts in the semantic group DISORDER [9]. Concepts recognized as DISORDER in the abstract title and first two sentences are ranked based on the frequency of occurrence.

The intervention extractor also generates a ranked list of interventions based on the semantic type information (for example, Therapeutic or Preventive Procedure, Clinical Drug, or Diagnostic Procedure) and positional information. Concepts frequently occurring in the titles, aims, and methods sections of an abstract are ranked higher. In unstructured abstracts, concepts extracted from the first third of the abstract are favored.

The top ranking DISORDER (or more, in case of a tie) and intervention(s) are extracted as the focus of the study.

The extracted problems and interventions were evaluated using the official MeSH indexing for the documents, which is not used by the extraction modules. We computed recall and precision for the problem extractor using starred MeSH descriptors of the DISORDER type for each document and for the intervention extractor using all starred intervention type headings as reference standards.

RESULTS

The results of the MTI experiment are provided in Table 1. The table includes descriptive information at the beginning as well as three sections of results: the overall results followed by those for citations with only a title and then those with both title and abstract. Table 2 contains the results for the extraction experiment with a section for each of problem and intervention extraction.

In Table 1 we see that the MTI results are much better for citations with an abstract. Also, the conservative views consistently outperform the aggressive views, with AllFilter scoring best. Furthermore, while the aggressive views exhibit the normal tradeoff between precision and recall for a given experiment, the conservative views violate that norm, i.e., both recall and precision improve as more filtering is done on the conservative views. This shows that the filtering is finding more results without introducing noise.

	Base	AutoFilter	AllFilter	AggrBackoff	Aggressive
Citations	9,998	9,999	9,999	9,999	9,998
Indexed MHs	115,877	115,877	115,877	115,877	115,877
MTI Recommendations	188,531	188,004	187,491	181,376	180,501
All Citations					
Correct MTI Recommendations	58,390	58,538	58,584	56,417	56,324
% of Indexed MHs **(Recall)**	50.40%	50.52%	50.56%	48.69%	48.61%
% of MTI Recommendations **(Precision)**	30.97%	31.14%	31.25%	31.10%	31.20%
F_2	41.68%	41.84%	41.92%	40.97%	40.99%
Title-Only Citations					
Correct MTI Recommendations	2,728	2,741	2,750	2,582	2,559
% of Indexed MHs **(Recall)**	19.69%	19.78%	19.85%	18.64%	18.49%
% of MTI Recommendations **(Precision)**	43.77%	44.29%	44.69%	44.04%	44.15%
F_2	24.11%	24.26%	24.36%	23.07%	22.93%
Title/Abstract Citations					
Correct MTI Recommendations	55,662	55,797	55,834	53,835	53,765
% of Indexed MHs **(Recall)**	54.57%	54.69%	54.73%	52.77%	52.70%
% of MTI Recommendations **(Precision)**	30.53%	30.69%	30.79%	30.67%	30.77%
F_2	43.23%	43.38%	43.46%	42.55%	42.59%

Table 1: MTI results

	Base	AutoFilter	AllFilter	AggrBackoff	Aggressive
Problem extraction					
Citations	5,262	4,601	4,148	3,824	3,377
Recall	33.05%	37.50%	39.89%	38.97%	40.66%
Precision	23.66%	30.62%	32.59%	30.57%	33.69%
F_2	29.19%	34.89%	37.17%	35.70%	38.04%
Intervention extraction					
Citations	6,695	6,342	5,959	4,214	4,172
Recall	23.57%	29.14%	30.24%	32.76%	33.20%
Precision	15.41%	17.07%	18.76%	19.16%	19.10%
F_2	20.03%	23.58%	25.12%	26.49%	26.64%

Table 2: Extraction results

From Table 2, on the other hand, we see that the aggressive views generally outperform the conservative views.

We used a two-tail paired t-test to determine if the differences observed in MTI and problem and intervention extraction based on five views are statistically significant. All derived views are significantly better suited than the Base view for all experiments except for MTI with the aggressive views. In fact the only cases in which results for different views were not statistically significant are:

- MTI and Intervention extraction with AggrBackoff vs. Aggressive; and

- Problem and Intervention extraction with AutoFilter vs. AllFilter.

The lack of statistical significance between the AggrBackoff and Aggressive views is not surprising since they are only slightly different anyway.

Our precision values for the problem and intervention extraction are much lower than those obtained in our previous evaluations. This seeming drop in performance is caused by the fact that in the CQA system, non-clinical publications and publications without abstracts are filtered out prior to the extraction. Moreover, the previous evaluation was conducted on clinical articles retrieved using a PubMed search for a starred MeSH heading for several disorders. Because the focus of this paper is the influence of the views on our processing, we chose to conduct the evaluation on the same random sample to make the influence of the view comparable for the two applications.

A possible explanation for the drop in performance for problem extraction with aggressive backoff is that the extra ambiguity introduced overwhelms any potential benefit.

DISCUSSION

The results obtained for the MTI experiments are consistent with what we expected. However, the corresponding results for extraction were more complex. In order to provide some perspective on the latter results, we show an example of extraction over the various views using Citation 17680185 from the LNCV document collection (see Figure 1):

The gold standard (GS) for this example consists of

TI - Effects of spinal anesthesia on the peripheral and deep core temperature in elderly diabetic patients undergoing urological surgery.

MH - Aged
MH - Aged, 80 and over
MH - ***Anesthesia, Spinal**
MH - Body Temperature/*physiology
MH - **Diabetes Mellitus/*physiopathology**
MH - Foot/physiology
MH - Forehead/physiology
MH - Humans
MH - Intraoperative Period
MH - Male
MH - Middle Aged
MH - Time Factors
MH - ***Urologic Surgical Procedures, Male**

Figure 1: Excerpt from citation 17680185

the problem *Diabetes Mellitus/*physiopathology* and the interventions **Anesthesia, Spinal* and **Urologic Surgical Procedures, Male*, all highlighted in bold in Figure 1.

All views correctly identified the problem and failed to identify one of the interventions (see Table 3 below).

The Base view mapped the text *body temperature* (from the abstract) to the diagnostic procedure *Body temperature measurement*, and it mapped the text *18 male patients (ASA physical status I or II)* to Aspirin (because of *ASA*). It consequently suggested these concepts as interventions.

The AllFilter view no longer extracted *Body temperature measurement* as an intervention; therefore *Spinal Anesthesia* moved to the top of the list.

Finally, although the Aggressive method suggested *Urologic Surgical Procedures*, it did not match the more specific concept *Urologic Surgical Procedures, Male* in the gold standard.

CONCLUSION

Our experiments indicate that conservative view construction is best suited to literature applications but that aggressive views might be useful for clinical applications. This will presumably be even more true as further backoff strategies are discovered to

improve the aggressive view. Combining the conservative and aggressive approaches is likely to prove useful, too. We will apply these observations as we continue to refine our literature UMLS content view and as we embark on the even more ambitious task of constructing multiple clinical content views.

View	#extracted problems	Extracted main problem(s)	GS problem	#extracted interventions	Extracted main intervention(s)	GS interventions
Base	8	Diabetes Mellitus	Diabetes Mellitus	8	1. Aspirin 2. Body temperature measurement	1. Spinal Anesthesia 2. Urologic Surgical Procedures, Male
AutoFilter	7	Diabetes Mellitus		5	1. Body temperature measurement 2. Spinal Anesthesia	
AllFilter	5	Diabetes Mellitus		4	1. Spinal Anesthesia	
Aggr-Backoff	6	Diabetes Mellitus		3	1. Spinal Anesthesia 2. Urologic Surgical Procedures	
Aggressive	6	Diabetes Mellitus		3	1. Spinal Anesthesia 2. Urologic Surgical Procedures	

Table 3: Extraction example results

Acknowledgements

The authors greatly appreciate the contribution to LNCV meeting discussions that led to this research from our colleagues Tom Rindflesch, Suresh Srinivasan, Kin Wah Fung, Allen Browne and Graciela Rosemblat.

This study was supported in part by the Intramural Research Programs of the National Institutes of Health, National Library of Medicine. Aurélie Névéol was supported by an appointment to the National Library of Medicine Research Participation Program administered by the Oak Ridge Institute for Science and Education through an inter-agency agreement between the U.S. Department of Energy and the National Library of Medicine.

References

[1] *Unified Medical Language System (UMLS) Documentation*. National Library of Medicine, Bethesda, MD. http://www.nlm.nih.gov/research/umls/documentation.html.

[2] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*. 2001;:17-21.

[3] Rogers WJ and Aronson AR. *Filtering the UMLS Metathesaurus for MetaMap: 2007 Edition*. National Library of Medicine, Bethesda, MD. http://skr.nlm.nih.gov/papers/references/filtering07.pdf.

[4] Shooshan SE and Aronson AR. *Ambiguity in the UMLS Metathesaurus: 2007 Edition*. National Library of Medicine, Bethesda, MD. http://skr.nlm.nih.gov/papers/references/ambiguity07.pdf.

[5] *Unified Medical Language System: Preface to the 2005AA Documentation*. National Library of Medicine, Bethesda, MD. http://www.nlm.nih.gov/research/umls/archive/2005AA/umlsdoc_preface.html.

[6] Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, Rindflesch TC, Wilbur WJ. The NLM Indexing Initiative. *Proc AMIA Symp*. 2000;:17-21.

[7] Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. *Medinfo* 2004;11(Pt 1):268-72.

[8] Demner-Fushman D, Lin J. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics* 2007;33(1): 63-104.

[9] McCray AT, Burgun A, and Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. *Medinfo* 2001;10(Pt 1): 216–220.