NLM's I2b2 Tool System Description

James G. Mork[1], MSc, Olivier Bodenreider[1], MD, PhD,
Dina Demner-Fushman[1], MD, PhD, Rezarta I. Dogan[2], PhD, François-Michel Lang[1], MSE,
Zhiyong Lu[2], PhD, Aurélie Névéol[2], PhD, Lee Peters[1], MSc,
Sonya E. Shooshan[1], MLS, Alan R. Aronson[1], PhD

[1]Lister Hill National Center for Biomedical Communications (LHNCBC),
[2]National Center for Biotechnology Information (NCBI),
U.S. National Library of Medicine, Bethesda, National Institutes of Health, MD 20894

Abstract

The i2b2 medication extraction challenge provided us with an opportunity to assess the usability of publicly available drug-related resources on clinical text and to contribute to the generation of a publicly available collection of annotated clinical notes. The challenge also presented us with a chance to evaluate how MetaMap, our UMLS concept recognition tool, would work on discharge summaries and to roll the knowledge gained back into MetaMap development. Our approach to identify drug-related entities within the scope of this challenge relied on the use of look-up lists and rules built solely with publicly available resources. Preliminary results show promise with the clinical drug information specific entity lists. However, more sophisticated methods will be needed to improve the identification of the reason and duration elements of drug mentions.

INTRODUCTION

The Lister Hill Natural Language Processing (NLP) Content View (LNCV) project [1] has shown that creating a domain specific subset of the Unified Medical Language System® (UMLS®) Metathesaurus® can improve recognition of clinical text via NLP tools. The i2b2 medication extraction challenge (referred to as just 'challenge' for the rest of the paper) provided us with a practical means of extending this earlier work by looking at a different form of clinical text as well as the opportunity to develop a set of drug-specific lookup lists and identification rules that might be incorporated into our suite of NLP applications.

We developed a straightforward tool that relied heavily on a set of lookup lists to identify the drugs and their components (mode, dosage, duration, and frequency); we then utilized a combination of MetaMap [2] and a lookup list from the Regenstrief Institute for Health Care and the Department of Medicine Gopher order entry system [3] to identify what reasons, if any, were associated with each of the drug occurrences.

SYSTEM DESCRIPTION

Lookup list development

The first step in the process of acquiring lookup lists of terms relevant to medications was identifying the publicly available resources we were going to use. Although many of the resources have items in common, each of the resources was added for specific reasons. Figure 1 graphically depicts where the data came from with arrows connecting the sources and the lists where they made contributions.

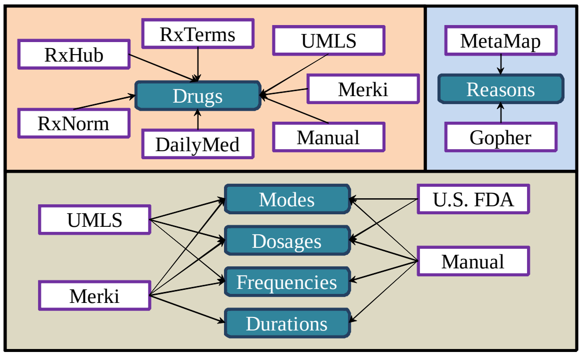


Figure 1: Lookup Lists and Their Sources

The drug identification list was created using DailyMed [4] for a list of common prescription drug names. We then added RxTerms [5], Ingredients and Brand Names from RxNorm [6], and a list of drugs from Merki [7] for a comprehensive list of drugs and their component ingredient names. We added pharmacologic classes (e.g., *vasopressors*) by extracting from the UMLS all the descendants of high-level concepts for pharmacologic preparations and added a list of classes from Merki. RxHub [8] provided us with a list of common drug name misspellings. The U.S. Food and Drug Administration (FDA) Structured Product Labeling web site [9] provided us with extensive lists of Dosage Forms and Routes of Administration. Specific filtering of the UMLS and additional lists from Merki also provided information for the dosage, modes, and frequencies lists. Merki was also used to create the duration list. Finally, manual curation was done to extend all of the lists based on reviews of the tool results for the training collection. For this last step, we specifically looked at the "missed" or not used tokens for each of the lines and assigned the text to the lists as appropriate.

Reason identification

We used both MetaMap and a lookup list derived from the Gopher system to identify reasons for prescribing drugs in this challenge.

MetaMap was designed to identify UMLS Metathesaurus concepts in biomedical text and does a very good job of this for well behaved text. For this challenge we investigated some new uses for MetaMap, but ended up only using MetaMap to identify the reasons for prescribing a drug. To restrict MetaMap to just looking for reasons, we limited MetaMap to only using the twelve Semantic Types from the *Disorders* Semantic Group [10], and because of the type of text we were dealing with in this challenge we included the *Clinical Attribute* Semantic Type as well.

In this challenge, the discharge summaries sometimes had misspellings, acronyms/abbreviations, and different ways of stating a medical reason for prescribing a drug. While MetaMap was able to identify some of the spelling variations and any text inversions, it was limited to the contents of the UMLS Metathesaurus. The Gopher lookup list was introduced to expand our capabilities and to assist with these less well behaved occurrences. The Gopher list was derived from menu items in the order entry system and represents names, aliases, and synonyms for diagnoses, procedures, tests, and drugs. We specifically used the names and synonyms for the challenge.

Section identification

Identifying the sections within the discharge summaries allowed us to pick which sections we wanted to process and assisted us in limiting the scope of combining drugs, reasons, and components. For example, we did not want to process sections that discussed the patient's allergies because of the guidelines for this challenge. Another reason for ignoring specific sections was to try and eliminate personal names from triggering reasons (e.g., first name *Brock* from discharge summary 236076 triggers UMLS concept *Middle Lobe Syndrome* which is the MeSH® Main Heading for *Brock's Syndrome*).

Processing

The diagram in Figure 2 details the straightforward processing our tool performed on each of the discharge summaries for this challenge.

1. Read the i2b2 discharge summaries into the tool
2. Each line tokenized, section and drug list boundaries identified

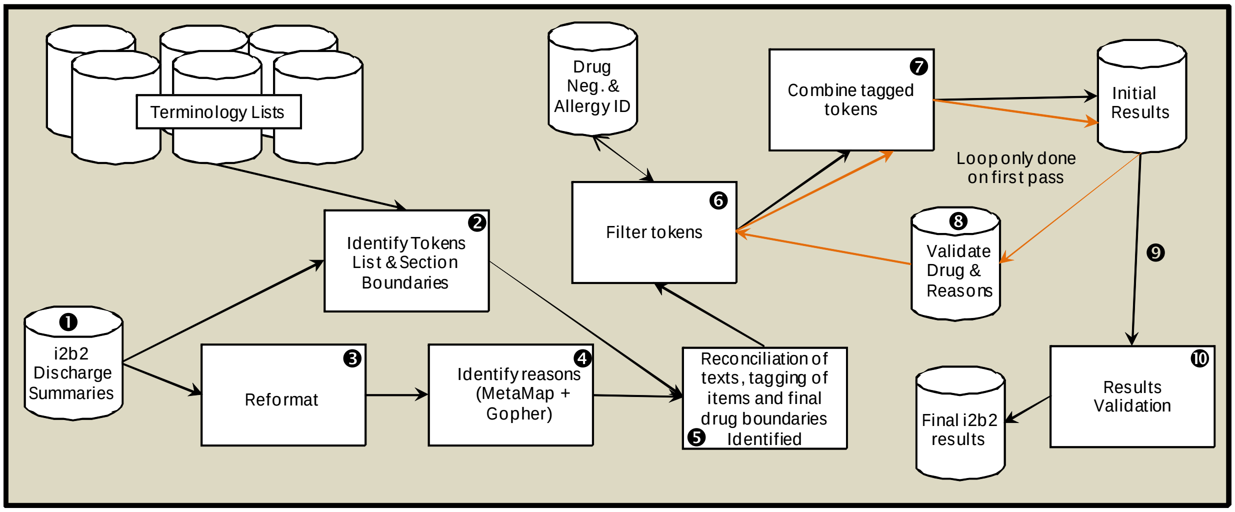

Terminology Lists
Drug Neg. & Allergy ID
❼ Combine tagged tokens
Initial Results
Loop only done on first pass
❷ Identify Tokens List & Section Boundaries
❻ Filter tokens
❽ Validate Drug & Reasons
❾
❶ i2b2 Discharge Summaries
❸ Reformat
❹ Identify reasons (MetaMap + Gopher)
Reconciliation of texts, tagging of items and final drug boundaries ❺ Identified
Final i2b2 results
❿ Results Validation

Figure 2: NLM's i2b2 Tool Processing Flow

3. Text was reformatted to ASCII MEDLINE format for MetaMap processing and sections we did not want to process were not included

4. MetaMap processing and Gopher list reason identification

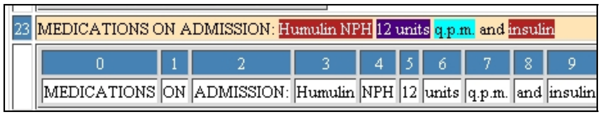
5. Reason locations were reconciled with the original summary text; and component tagging to identify drugs, modes, dosages, durations, and frequencies and drug boundaries were marked.

6. Filtering to add, remove, and extend tagged items. Filtering involved simple rules and a "bad drugs" list for what should be removed (e.g., *insulin* within "*insulin*-dependent diabetes". We also had rules for limiting the scope of a drug to try and eliminate the crossover of components. We also developed a program to identify negated and allergy specific drugs (e.g., *should not take aspirin*) to remove false positives.

7. Matching up drug names to components and reasons. We had a small set of rules for combining drugs and rules – for example, if we found <*drug*> *for* <*reason*> in the text, we would combine the two.

8. We also developed a rule-based program to identify valid pairings of drugs and reasons via a constrained traversal of the UMLS relations. The validity checking program was used to positively identify valid combinations, but not for removing any combinations. An example of a valid combination would be *albuterol* and *asthma* where there is a direct link between an ingredient and a disease.

9. A set of five files were created for each discharge summary as a result of our tool processing – most of these were for debugging purposes: 1) a file with just the i2b2 formatted results, 2) a file with the drug/reason pairings for our validation program in step 8 above, 3) a file showing all of the untagged text in context for each line in each discharge summary used for almost all of our manual curation efforts, 4) a detailed HTML file with color coded text depicting the final decisions showing which tagged tokens (mode, duration, dosage, frequency, and reason) where combined with which drugs, and 5) an informational HTML file also color coded only illustrating the raw tagging that was done by the tool. Figure 3 shows an example of this last informational HTML file for line 23 of discharge summary 23538. The image shows that our tool has identified *Humulin NPH* as a drug, a dosage of *12 units*, frequency of *q.p.m.*, and second drug *insulin*. Each line is repeated on a second line where each of the tokens is identified and numbered according to the challenge rules for tokenization. Most of our team viewed this web page for each discharge summary while annotating it in the first round.



23 MEDICATIONS ON ADMISSION: Humulin NPH 12 units q.p.m. and insulin

0	1	2	3	4	5	6	7	8	9
MEDICATIONS	ON	ADMISSION:	Humulin	NPH	12	units	q.p.m.	and	insulin

Figure 3: Tool Information View Example (23538)

10. Validation is done at the end to verify compliance with the challenge requirements.

References

[1] Demner-Fushman D, Mork JG, Shooshan SE, Aronson AR, UMLS Content Views Appropriate for NLP Processing of the Biomedical Literature vs. Clinical Text. Accepted for AMIA 2009.

[2] Aronson AR, Lang FM. The Evolution of MetaMap, a Concept Search Program for Biomedical Text. Accepted for AMIA 2009.

[3] McDonald CJ, Tierney WM: The medical gopher- A microcomputer system to help find, organize and decide about patient data. *West J Med* 1986 Dec; 145(6):823-829.

[4] http://dailymed.nlm.nih.gov

[5] http://wwwcf.nlm.nih.gov/umlslicense/rxtermApp/rxTerm.cfm

[6] http://www.nlm.nih.gov/research/umls/rxnorm/

[7] Gold S, Elhadad N, Zhu X, Cimino JJ, Hripcsak G. Extracting structured medication event information from discharge summaries. AMIA Annu Symp Proc. 2008 Nov 6:237-41.

[8] Fung KW, Applied Medical Terminology Research. A Report to the Board of Scientific Counselors. April 2009. Page 32. http://www.lhncbc.nlm.nih.gov/lhc/docs/reports/2009/tr2009001.pdf#page=33

[9] http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/ucm162038.htm.

[10] McCray AT, Burgun A, and Bodenreider O. Aggregating UMLS semantic types for reducting conceptual complexity. *Stud Health Technol Inform.* 2001;84(Pt 1):216-20.