Extracting Rx information from clinical narrative

James G Mork,[1] Olivier Bodenreider,[1] Dina Demner-Fushman,[1] Rezarta Islamaj Doğan,[2] François-Michel Lang,[1] Zhiyong Lu,[2] Aurélie Névéol,[2] Lee Peters,[1] Sonya E Shooshan,[1] Alan R Aronson[1]

► Additional appendix is published online only. To view this file please visit the journal online (http://jamia.bmj.com).

[1]Lister Hill National Center for Biomedical Communications (LHNCBC), US National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA
[2]National Center for Biotechnology Information (NCBI), US National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

Correspondence to
Dr Alan R Aronson, National Library of Medicine, Building 38A, Room 9N-905, 8600 Rockville Pike, MSC-3826, Bethesda, MD 20894, USA; alan@nlm.nih.gov

Received 23 February 2010
Accepted 25 June 2010

ABSTRACT

Objective The authors used the i2b2 Medication Extraction Challenge to evaluate their entity extraction methods, contribute to the generation of a publicly available collection of annotated clinical notes, and start developing methods for ontology-based reasoning using structured information generated from the unstructured clinical narrative.

Design Extraction of salient features of medication orders from the text of de-identified hospital discharge summaries was addressed with a knowledge-based approach using simple rules and lookup lists. The entity recognition tool, MetaMap, was combined with dose, frequency, and duration modules specifically developed for the Challenge as well as a prototype module for reason identification.

Measurements Evaluation metrics and corresponding results were provided by the Challenge organizers.

Results The results indicate that robust rule-based tools achieve satisfactory results in extraction of simple elements of medication orders, but more sophisticated methods are needed for identification of reasons for the orders and durations.

Limitations Owing to the time constraints and nature of the Challenge, some obvious follow-on analysis has not been completed yet.

Conclusions The authors plan to integrate the new modules with MetaMap to enhance its accuracy. This integration effort will provide guidance in retargeting existing tools for better processing of clinical text.

INTRODUCTION

Extraction of the elements of medication orders from clinical narrative is a preliminary step in many important applications of medical informatics. These applications include but are not limited to: support of quality assurance through reconciliation of patient's medication lists and clinical notes[1 2]; detection of adverse reactions to drugs[3] and medication non-compliance[4]; study of a population's response to a drug[5]; support of care plan development[6]; and identification of inactive medications.[7]

Whereas evaluation of the individual efforts in extraction of medication names from biomedical literature could use 'found data', such as Medical Subject Headings (MeSH) assigned to MEDLINE abstracts in the manual indexing process,[8] until recently no annotated resources for evaluation of extraction of medication orders from clinical narrative were publicly available. The opportunity to evaluate our named entity extraction methods and to contribute to development of an annotated publicly available large collection of clinical notes presented itself with the third i2b2 (Informatics for Integrating Biology and the Bedside) Medical Extraction Challenge.[9]

To date, most algorithms and systems for extraction of drug order elements are knowledge-based. In fact, the absence of any large annotated collection makes it difficult to use supervised machine learning. In contrast, the availability of nomenclatures such as RxNorm[10] (which contains drug names, ingredients, strengths, and forms) encourages the use of rule-based systems. For example, Evans *et al*[11] developed a set of about 50 rules encoded as regular expressions to identify drug dosage objects and their attributes. A natural language processing (NLP) system augmented with the above rules and two lexicons (one containing drug names extracted from the Unified Medical Language System (UMLS)[12] and another one containing unusual words and abbreviations found in drug dosage phrases) identified about 80% of drug dosage expressions. Gold *et al*[1] expanded the definition of drug dosage of Evans *et al* and implemented a system (the MERKI parser) that uses an RxNorm-based lexicon to extract known drug names and contextual clues to extract out-of-vocabulary drug names. Xu *et al*[13] developed an approach that attempts to extract a formal medication model (consisting of the drug name, signature modifiers and temporal modifiers) from clinical text using a chart parser and a semantic grammar, and backs off to regular expressions if the chart parser fails.

The US National Library of Medicine (NLM) tool (referred to as NLM's i2b2 Challenge Tool or simply, the Tool) developed to extract all fields originally defined in the i2b2 medication extraction guidelines is also knowledge-based and relies on lexical–semantic processing and pattern matching similar to the above systems. Our approach differs from the previously explored ones in that we (1) expanded a large number of term lists obtained for each element of drug phrases generating potential spelling variants and mining the UMLS for related terms as well as using corpus-based expansion, (2) developed a module for identification of negated drug mentions, (3) applied a UMLS-based approach to identification of reasons for medication orders, and (4) developed a module for validating drug and reason combinations.

METHODS

Early in the planning phase for this Challenge, the decision was made to use simple rules and lookup lists of various entities because of the time constraints of the Challenge. Our processing of the

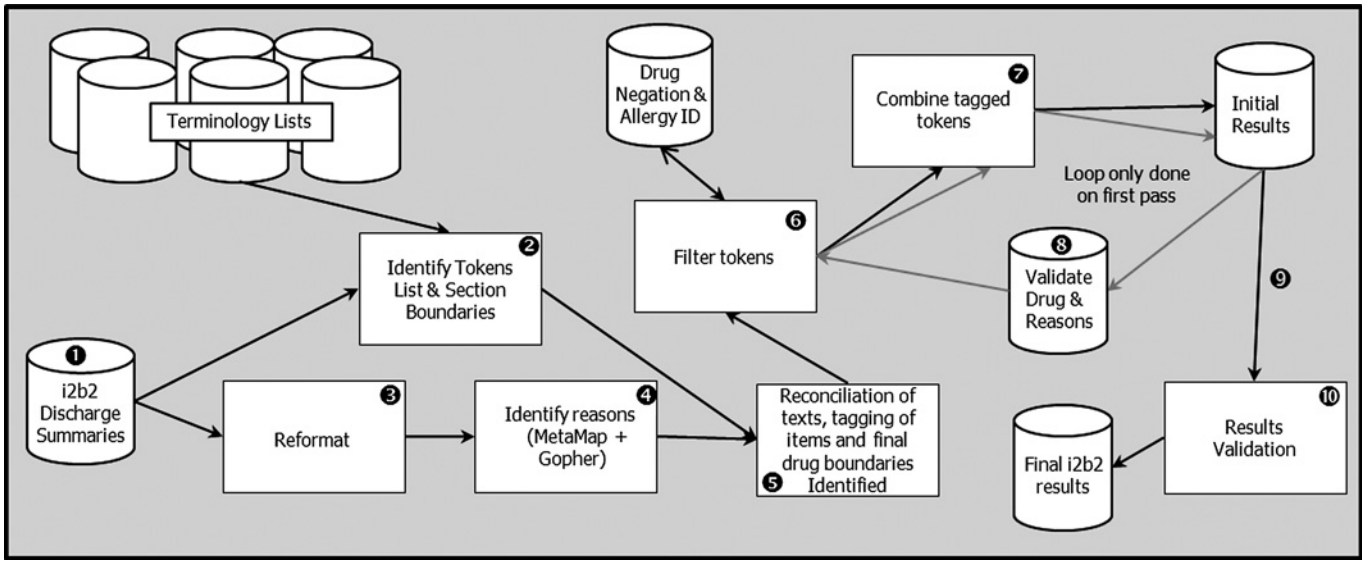

Figure 1 Processing flow diagram.

discharge summaries for this Challenge was relatively straightforward and is depicted in figure 1. This section follows the course of our processing efforts beginning with a description of the lookup lists developed for the Challenge. Complete details can be found in our appendix (full-length paper online at http://jamia.bmj.com).

Development of lookup lists required for the challenge

The discovery of coverage gaps in our terminology resources (eg, short forms of drug names such as 'aspart' are not always covered in the UMLS, although the long form, 'insulin aspart', maps to two concepts) led to the decision to augment our initial resources with lookup lists. The lists that we developed used existing, publicly available resources with some minor manual curation based on processing the training set and reviewing what was missed by the Tool described here. Although many of the resources have items in common, each of the resources was added for specific reasons. Figure 2 graphically depicts the data sources with arrows connecting the entities and the lists where they made contributions.

The drug identification list was created using DailyMed[14] for a list of common prescription drug names. We then added display names from RxTerms,[15] Ingredients and Brand Names from RxNorm, and a list of drugs, drug classes, dosages, modes, frequencies, and durations from MERKI. In an attempt to complement the list of drugs we already had, we started looking at pharmacologic classes (eg, diuretics), as opposed to drug names, and added about 5000 names from 1360 UMLS concepts. RxHub,[16] which is derived from drug names obtained from deidentified patient medication records, provided us with a list of common drug name misspellings. The US Food and Drug Administration (FDA) Structured Product Labeling website[17] provided us with extensive lists of dosage forms (dosages) and routes of administration (modes). Finally, manual curation was performed to extend all of the lists based on reviews of the Tool results for the training set.

Discharge summaries read into the program and tokenized

Each line was tokenized using white space as the token boundary. List boundaries were simply identified by which sections corresponded to the Challenge list of valid 'list' sections. Sentence boundaries were identified using the simple rule of finding a 'period' followed by spacing as long as the previous character was not a number. Sentence boundaries helped to define the extent of both drugs and reasons. Section identification was most crucial to this Challenge for several reasons: it (1) allowed us to decide if we wanted to process specific sections or ignore them, (2) assisted in limiting the scope of drugs and reasons, (3) was instrumental in determining whether a drug was in a 'list' or 'narrative', and (4) helped eliminate some ambiguity (eg, not identifying drugs within Allergy sections). Candidate section names were defined as all strings occurring at the beginning of a line, consisting of uppercase letters only (a mixed-case review was attempted, but found to be too noisy), and followed by a period, a colon, or the end of the line. We identified 10 454 such potential section names, 937 of them unique. The list of unique names was then manually reviewed, scrubbed, and some mixed-case section names were manually added to the list—for example, 'Attending'. We consequently created a list of 21 triggers (table 1) that denoted sections we could ignore. We ended with 632 section names extracted from the training set.

Text reformatted into a single text line

Early testing showed that by simply processing the summaries line by line, we ended up missing some drugs and reasons

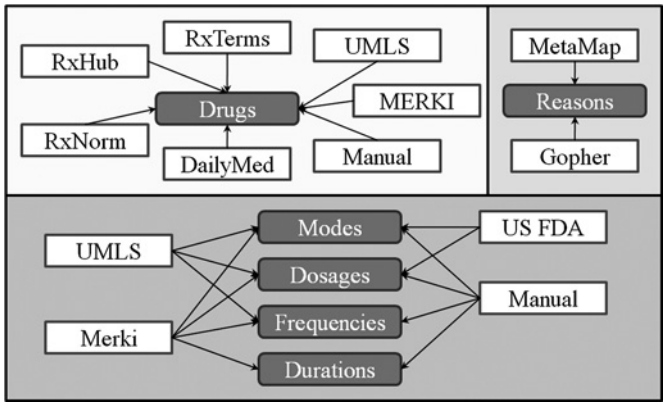

Figure 2 Lookup lists and their sources.

Table 1 List of trigger phrases for sections to be ignored

lab	laboratory	laboratories	allergies	allergy
attending	fam hx	family history	family hsitory	discharge date
service	labs	escription document	dictated by	entered by
vital sign	vitals	signs	vital signs	diet

because the text was broken across lines. So, once the sections were identified, we combined all of the text to be processed into a single line. A mapping between the reformatted and original text was maintained.

Reasons identified using MetaMap and exact matches from the Gopher list

We used both MetaMap[18] and a list derived from the Gopher[19] project to identify reasons. 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 most 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 coverage and to assist with less well-behaved occurrences. In the end, the two approaches seemed to complement each other fairly well. We also maintained a 'bad reason' list to eliminate as many false positives as possible (see Filtering section below).

Reasons reconciled with the original text and tagged using the mapping information from the single free-text line back to the original discharge summary

We used exact text matches to the lookup lists to tag drugs, modes, dosages, durations, and frequencies. Drug boundaries were also identified by noting the first position of each drug so we could know when we came to the end of the current drug during filtering. Drug boundaries expanded left and right depending on where the components were identified, with the final drug boundary encompassing the drug name and any of its associated components.

Filtering performed to add, remove, and extend tagged items

Filtering involved simple rules, a 'bad reason' trigger list (eg, 'ruled out for'), and a 'bad drugs' list for what should be removed (eg, 'insulin' within 'insulin-dependent diabetes'). We developed rules for limiting the scope of a drug to try to eliminate the crossover of components, and we also tried to identify non-active medications (eg, 'should not take aspirin') and allergy-specific drugs to remove false positives. Simple rules for expanding components by looking at the tokens to the left and right of the component were developed as needed.

Drug/reason pairings identified

Once drugs and reasons had been initially identified, we attempted to match each drug name with a nearby reason. Initially we had a very simple rule to use the closest reason if there were two possibilities. This was refined to ensure that reason assignment did not violate a drug, list, or section boundary. We also created a small set of trigger phrases to use in combining certain nearby reasons and drugs (figure 3). In some cases, we allowed multiple reasons for a drug if they were next to each other and connected with a comma, 'and', or 'or'.

Validation of drug/reason pairings

Once drug/reason pairings were identified, we attempted to validate the pairings via knowledge contained in the UMLS. The validation of the drug/reason pairings was accomplished via a constrained traversal of the UMLS relations involving two main steps as described below.

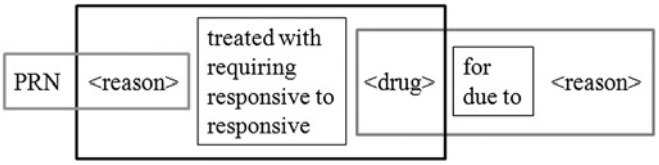

Figure 3 Simple reason grouping rules.

Drugs and reasons were first mapped to UMLS concepts, using exact and normalized matches, and further restricting mappings to the semantic group 'Chemicals & Drugs' and 'Disorders', respectively. All successful mappings were considered, including several pairs of UMLS concepts generated by one original drug/reason pairing.

Selected UMLS relations were then used to identify plausible relations between drugs and reasons. The key relations were provided by the NDF-RT source vocabulary where ingredients are associated with diseases through 'may_treat' and 'may_prevent' relationships.

The algorithm did not explore all paths, but rather stopped at the first path reached between the drug and the reason. For example, 'albuterol/asthma' was identified through a direct link between ingredient and disease. A total of 9415 possible drug/reason pairings were found, with 2785 of these having at least one path through the UMLS tying them together.

RESULTS

We finished fourth overall out of 20 teams that participated in the Challenge. Since two of the three teams who scored best had pre-existing systems that were modified for the Challenge, we were pleased that a system developed expressly for the Challenge performed so well. The lessons learned during this effort are being evaluated for inclusion in our NLP tool suite. Results are shown in table 3 and table 7 in the i2b2 *JAMIA* overview paper.[20] It is clear from table 7 that all teams had significant problems with identifying both durations and reasons.

DISCUSSION

In general, we are satisfied with our vocabulary and rule-based identification of drug names, doses, modes, and frequencies. The lack of significant difference between our exact and inexact scores confirms this view, as it shows that we either found the entire element or missed it completely. Our dose and duration results are satisfactory considering they are based on very simple heuristics. However, the approach is brittle in the presence of pattern changes in the middle of an enumeration of drugs. Deeper understanding of the context is needed to overcome this weakness.

Low scores for durations and reasons, on the other hand, show that our methods are clearly insufficient for those drug elements. In the absence of creating a full-fledged natural language understanding system, some improvement might be achieved using corpus-based methods. Any corpus-based methods would need to be judiciously applied given their known weaknesses: they are noisy if not supervised, and they are ambiguous even when supervised. For example, using our corpus-based expansion, we identified 'HCT' as an abbreviation of 'hydrochlorothiazide' (more commonly abbreviated as 'HCTZ'); however, 'HCT' is also common shorthand for 'hematocrit'.

Finally, we intend to incorporate some of our Tool's features into the MetaMap algorithm. Specifically, we will include the

overall identification of drug mentions with the expectation that it will reduce ambiguity because of the coordination of a drug's elements. In addition, augmenting MetaMap's negation algorithm with the drug-specific negation detection developed for the Challenge should be useful in applying it to clinical text.

LIMITATIONS

Many of the limitations of this research occurred because we are reporting on the development of an NLP application in the context of a time-sensitive Challenge rather than fundamental research. In-depth analysis that we would normally have carried out will be performed in the future. Examples of such analysis include determining the relative contributions to our results from the many knowledge sources we used, a similar analysis of the contributions of the filtering rules, and a study to determine an optimal balance between the knowledge sources and the rules. In addition, the relations identified between drugs and diseases from selected UMLS relations are not intended to be used as a reference set of relations reflecting therapeutic intent. Rather, we use constraints on the UMLS graph of relations in order to identify plausible drug/reason relations for the purpose of validating drug/reason pairings. Despite the presence of many false positives and false negatives, our algorithm proved useful in the context of this Challenge.

Funding This work was supported by the Intramural Research Program of the NIH, National Library of Medicine. The project described was supported in part by the i2b2 initiative, Award Number U54LM008748 from the National Library of Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Gold S,** Elhadad N, Zhu X, *et al*. Extracting structured medication event information from discharge summaries. *AMIA Annu Symp Proc* 2008:237–41.
2. **Cimino JJ,** Bright TJ, Li J. Medication reconciliation using natural language processing and controlled terminologies. *Stud Health Technol Inform* 2007;**129**:679–83.
3. **Friedman C.** Discovering novel adverse drug events using natural language processing and mining of the electronic health record. *Artificial intelligence in medicine, 12th conference on artificial intelligence in medicine, AIME 2009*. Verona, Italy, 2009. Proceedings 2009.
4. **Turchin A,** Wheeler HI, Labreche M, *et al*. Identification of documented medication non-adherence in physician notes. *AMIA Annu Symp Proc* 2008:732–6.
5. **Sirohi E,** Peissig P. Study of effect of drug lexicons on medication extraction from electronic medical records. *Pac Symp Biocomput* 2005:308–18.
6. **Demner-Fushman D,** Seckman C, Fisher C, *et al*. A prototype system to support evidence-based practice. *AMIA Annu Symp Proc* 2008:151–5.
7. **Breydo EM,** Chu JT, Turchin A. Identification of inactive medications in narrative medical text. *AMIA Annu Symp Proc* 2008:66–70.
8. **Aronson AR,** Mork JG, Névéol A, *et al*. Methodology for creating UMLS content views appropriate for biomedical natural language processing. *AMIA Annu Symp Proc* 2008:21–5.
9. Informatics for integrating biology and the bedside, i2b2, a National Center for Biomedical Computing, third shared-task challenge in natural language processing for clinical data medication extraction challenge. https://www.i2b2.org/NLP/Medication.
10. **RxNorm.** http://www.nlm.nih.gov/research/umls/rxnorm.
11. **Evans DA,** Brownlow ND, Hersh WR, *et al*. Automating concept identification in the electronic medical record: an experiment in extracting dosage information. *Proc AMIA Annu Fall Symp* 1996:388–92.
12. **UMLS Knowledge Sources.** http://www.nlm.nih.gov/research/umls.
13. **Xu H,** Stenner SP, Doan S, *et al*. MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc* 2010;**17**:19–24.
14. **DailyMed.** http://dailymed.nlm.nih.gov.
15. **RxTerm.** http://wwwcf.nlm.nih.gov/umlslicense/rxtermApp/rxTerm.cfm.
16. **Fung KW.** Applied medical terminology research. A report to the board of scientific counselors. The RxHub Project—a sneak preview. April 2009:32. http://www.lhncbc.nlm.nih.gov/lhc/docs/reports/2009/tr2009001.pdf#page=33.
17. **FDA Structured Product Labeling Resources.** http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling.
18. **Aronson AR,** Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010;**17**:229–36.
19. **McDonald CJ,** Tierney WM. The medical gopher—A microcomputer system to help find, organize and decide about patient data. *West J Med* 1986;**145**:823–9.
20. **Uzuner Ö,** Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc* 2010;**17**:514–8.

Extracting Rx information from clinical narrative

James G Mork, Olivier Bodenreider, Dina Demner-Fushman, et al.

JAMIA 2010 17: 536-539
doi: 10.1136/jamia.2010.003970