Linguistic approach for identification of medication names and related information in clinical narratives

Thierry Hamon and Natalia Grabar

JAMIA 2010 17: 549-554
doi: 10.1136/jamia.2010.004036

Linguistic approach for identification of medication names and related information in clinical narratives

Thierry Hamon,[1] Natalia Grabar[2,3]

[1]LIM&BIO (EA3969), UFR SMBH Léonard de Vinci, Université Paris 13, 93017 Bobigny Cedex, France
[2]Centre de Recherche des Cordeliers, Université Pierre et Marie Curie - Paris6, UMR_S 872-20, Paris F-75006; Université Paris Descartes, UMR_S 872-20, Paris, F-75006; INSERM, U872-20, Paris F-75006, France
[3]HEGP AP-HP, 20 rue Leblanc, Paris F-75015, France

Correspondence to
Dr Thierry Hamon, Laboratoire d'Informatique Médicale et Bio-Informatique -Université Paris Nord, 74 rue Marcel Cachin, Bobigny Cedex 93017, France;
thierry.hamon@univ-paris13.fr

Received 24 February 2010
Accepted 29 June 2010

ABSTRACT

Background Pharmacotherapy is an integral part of any medical care process and plays an important role in the medical history of most patients. Information on medication is crucial for several tasks such as pharmacovigilance, medical decision or biomedical research.

Objectives Within a narrative text, medication-related information can be buried within other non-relevant data. Specific methods, such as those provided by text mining, must be designed for accessing them, and this is the objective of this study.

Methods The authors designed a system for analyzing narrative clinical documents to extract from them medication occurrences and medication-related information. The system also attempts to deduce medications not covered by the dictionaries used.

Results Results provided by the system were evaluated within the framework of the I2B2 NLP challenge held in 2009. The system achieved an F-measure of 0.78 and ranked 7th out of 20 participating teams (the highest F-measure was 0.86). The system provided good results for the annotation and extraction of medication names, their frequency, dosage and mode of administration (F-measure over 0.81), while information on duration and reasons is poorly annotated and extracted (F-measure 0.36 and 0.29, respectively). The performance of the system was stable between the training and test sets.

INTRODUCTION

Pharmacotherapy is an integral part of any medical care process and plays an important role in the medical history of patients. Acquiring accurate medication-related data is an important task. It is useful for improving patient safety and the quality of individual healthcare. Thus, pharmacovigilance[1 2] aims to prevent adverse drug effects. Medical[3–6] and pharmacological[7] decision systems are oriented towards prescription assistance: they improve medication reconciliation and reduce errors caused by misinterpretation of handwritten orders, incorrect doses, etc. With translational medicine, a better connection between clinical healthcare and biomedical research is established,[8 9] while the scientific literature helps biologists carrying out research on new drugs.[8 10] Knowledge about drugs is thus necessary, and medication-related information (eg, dosage, mode, time) provides even more precise knowledge.

Large-scale observation of data is necessary and becomes possible through extensive study of scientific literature and patient records. For this, structured data on prescriptions can be exploited,[11 12] but it has been observed that this type of data is often incomplete or out of date[13–15] and limited to prescriptions filled at a given hospital, and not at other places. Nevertheless, in scientific literature and clinical records, information on medication is buried in a mass of narrative text. To avoid this information becoming lost, we need specific tools and methods to detect, extract and exploit it.

BACKGROUND

Natural language processing (NLP) and text mining tools allow us to access relevant information within narrative documents. They perform parsing and analysis of unstructured documents in order to localize the data searched for. For instance, medication-related information may consist of a drug name, dose, frequency, duration, status and mode of administration. Detection of medication names is mostly dictionary-based: a nomenclature of drugs is used and their occurrences are detected in biomedical literature[16–18] or in clinical records.[18–22] It has been observed that the quality of such nomenclatures must be controlled,[19] as it has a direct impact on the quality of results. Approximate matching was proposed as a method of drug name recognition[20] and shown to improve extraction results compared with dictionary-based exact matching. Other methods aim to identify new drug names through naming conventions[23 24] or contextual rules.[25] Previous work has also addressed the extraction of drug-related information. The first study of this kind[29] focused on extracting drug names, a process improved by considering their context: dosage information allowed disambiguation of medication names. Extraction of drug-related data was also considered separately by research[26 30 31 33] and commercial[27 28] systems. The performance of these systems ranges from F-measures of 0.27 to 0.90 depending on the category of data: they are difficult to compare, as no common 'gold standard' has been used. Notice that applying such methods to database entries[32] significantly improves results (up to F-measure of 0.98). Common difficulties are related to incompleteness of drug lexica[19 20 26] and ambiguous drug names.[19 27 28]

RESEARCH QUESTIONS

In this work, we proposed to extract medication names and medication-related information, such as those underlined in the excerpt from box 1, from narrative discharge summaries. We proposed to go beyond the state-of-the-art and to address the following problems: (1) recognizing new medication names; (2) disambiguating medication names; (3) detecting contexts where drug names do not correspond to prescriptions.

Box 1 Excerpt from a narrative discharge summary with medication-related information (underlined) to be extracted

The patient is currently off diuretics at this time. Daily weights should be checked and if her weight increases by more than 3 pounds Dr Bockoven should be notified. The patient was also started on calcitriol given elevation of parathyroid hormone. Cardiovascular: Rate and rhythm: The patient has a history of atrial fibrillation with a slow ventricular response. The patient was started on metoprolol 12.5 mg p.o. q.6 h. for rate control, however, this dose was decreased to 12.5 mg p.o. twice a day, given some bradycardia on her telemetry. The patient was also started on Flecainide 75 mg p.o. q.12 h. She will continue on these two medications upon discharge.

We also evaluate our results through the common framework of the I2B2 NLP medication challenge held in 2009. This framework allows comparison between several automatic systems and NLP methods. We consider the categories targeted by the challenge (table 1): dosage, frequency, duration, mode of administration and reason for prescription, as well as the semantic relations between them. The NLP system designed exploits nomenclatures and terminologies, contextual rules and shallow parsing. Concurrent annotations may be proposed for a given token and then disambiguated.

COLLECTING AND PREPARING THE MATERIAL

Discharge summaries

Discharge summaries were provided by Partners Healthcare: they were written in English and were prepared and deidentified.[34] A total of 1249 documents were used, split into training (n=696) and test (n=553) sets. Within the training set, only 17 documents were manually annotated and provided as an illustration of annotation guidelines.

Terminologies and nomenclatures

We used two types of resource for the annotation (a total of 290 243 entries): drug nomenclature and pathology terms.

We created a medication list containing 243 869 entries mainly provided by RxNorm.[35 36] This list has three main limitations: the entries can be composed entries, common English words are used, and it is not exhaustive. To address the first two limitations, entries were split and cleaned up to remove ones such as 'golden eye', 'ginger', 'bermuda', 'vital', 'Marihuana' or 'water'. As for the third limitation, the list was enriched with drug names found in the training set. Moreover, we used therapeutic classes and groups of medications, found on the CDC website (http://www.cdc.gov/nchs/data/nhanes/nhanes_01_02/rxq_rx_b_doc.pdf). In addition, among the drug names, we distinguished 108 ambiguous entries that also referred to biological characteristics of patients (eg, 'red blood cells', 'magnesium', 'iron'). They were assigned a specific status.

Snomed International[37] proved to be an efficient and user-friendly source for NLP processing[38]; we used the 45 898 terms from the Diagnosis and Morphology axes for the detection of reasons. A total of 476 terms corresponding to patient problems in the training set were added to this resource.

Table 1 Examples of the targeted categories of information on drugs, as extracted from the excerpt given in box 1 (except for the values of the duration category)

Type	Abbreviation	Examples
Drug name	m	Calcitriol, metoprolol, flecainide, these two medications
Dosage	do	12.5 mg, 75 mg
Frequency	f	q.6 h, twice a day, q.12 h
Mode	mo	p.o.
Duration	du	7-day course, ×5 days, # for 7 days, 5 more days, 4 days
Reason	r	Elevation of parathyroid hormone, rate control

Negation markers

We exploited NegEx (http://www.dbmi.pitt.edu/chapman/NegEx.html) to detect negation and reduce the number of false positives. Negation markers consist of pre-negation (eg, 'deny', 'cannot', 'without') and post-negation (eg, 'free', 'was ruled out'). Some additional markers were added, making a total of 284 markers.

METHOD

Given the very small number of annotated documents available for tuning the systems (n=17), we used a rule-based approach: learning algorithms would require a larger training set. The system designed performs information extraction by three main steps: pre-processing, processing and post-processing (figure 1). The processing step is built on the Ogmios platform[39] suitable for the processing and annotation of large datasets and tunable to specialized areas. For pre- and post-processing steps, we developed specific modules to disambiguate and select the relevant annotations, to compute semantic relations, etc.

Pre-processing step

Input discharge summaries are full-text documents. To prepare them for the NLP tools, we first attempted to split them into sections and lists through the use of specific parsers and section markers (eg, 'discharge meds', 'history of present illness', 'family history', 'physical examination'). As these markers were not standardized across the discharge summaries, we supplemented them with contextual heuristics (eg, 'uppercase characters', 'punctuation'). Contextual heuristics were also used for the detection of lists and enumerations. Documents were then converted into XML format, with section and list tags. This step also computed offset data (number of lines and tokens) for the generation of the I2B2 output.

Processing step

The processing step was dedicated to linguistic and semantic annotation: we assigned semantic categories to textual entities and provided their semantic contexts. Our system supports concurrent annotations, while semantic contexts allow performance of their disambiguation. The annotation process was performed through the following main modules:

- The named entity recognizer (NER) identified frequency, dosage, duration and mode of administration. For this, specific automata were implemented as regular expressions (box 2). Preliminary disambiguation was performed in order to (1) select the longest match and avoid multiple annotations within nested strings (eg, 'ten minutes' was recognized as both frequency and duration entities), and (2) merge adjacent named entities of the same semantic type: 'q6h' and 'prn' were first recognized individually as frequency and then merged.
- Word and sentence segmentation was then performed. Having this step after the NER module allows the disambiguation of characters, such as punctuation, dashes, slashes, etc, that are widely used within discharge summaries often altering the segmentation process.

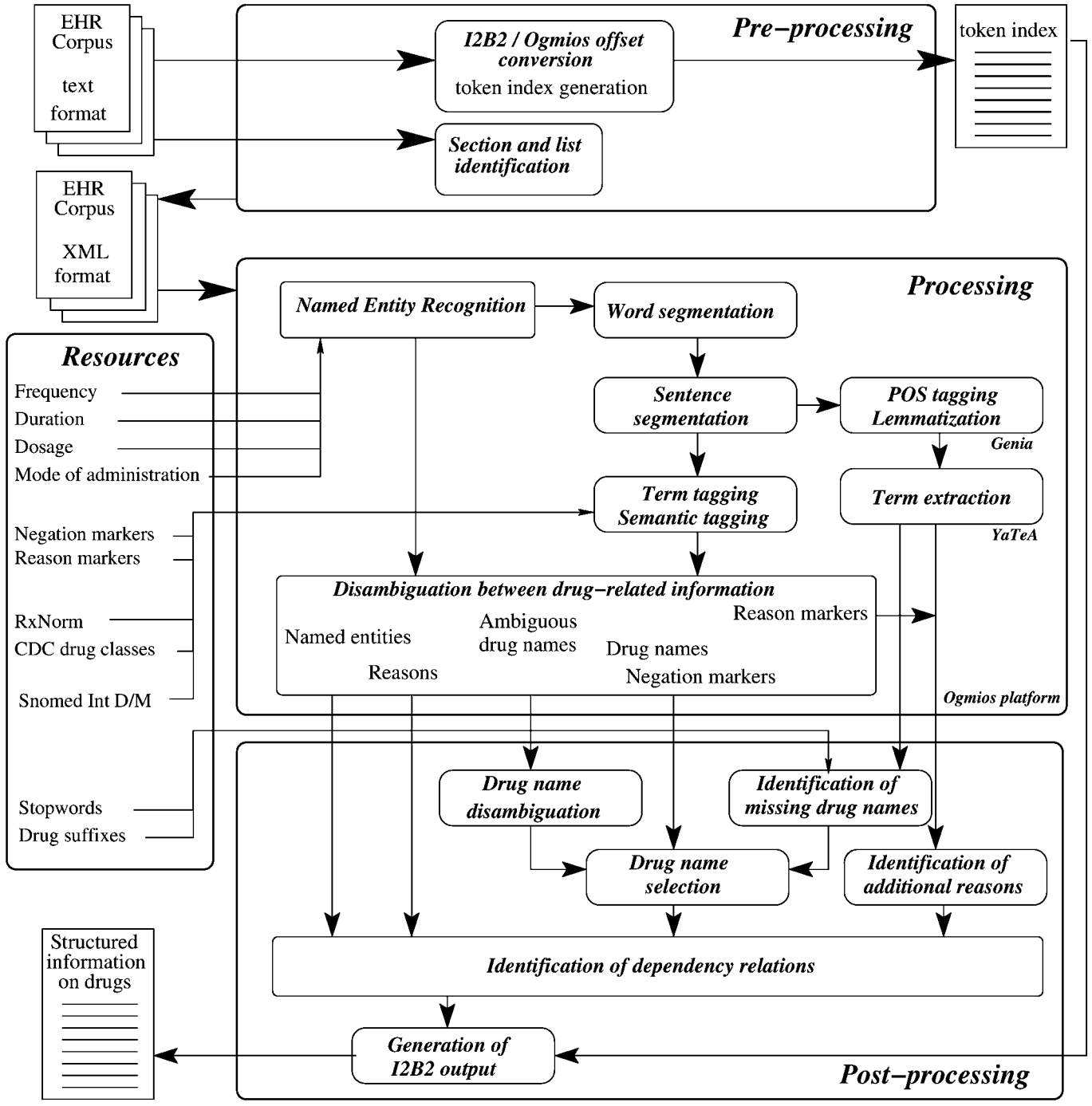

Figure 1 System architecture for the extraction of medication-related information and for establishing dependencies among the annotations. POS, part-of-speech.

- Term and semantic tagging was used to detect drugs and reasons. The system also performed the longest match and merged adjacent medication terms: in 'singulair (montelukast)', the two drugs correspond to two separate entities in our drug nomenclature.
- Term extraction was performed with $Y_A T_E A$[40]: it organizes the identification of missing medication names and reasons during the post-processing step. Part-of-speech tagging and lemmatization were performed with Genia.[41]

Post-processing step

In charge of several treatments on drugs and related information and computing dependency relations, the post-processing step exploits annotations from the processing step.

Disambiguation of medication names

Some medication names (eg, iron) are ambiguous: they can correspond to biological characteristics or drugs. They were first assigned a specific semantic tag. Then, if they occurred in listings or medication-related sections (box 3, example ii), their tags were modified into drug names. Otherwise, they were not considered (box 3, example i).

Detection of negative contexts and allergies

Our system deals with several contexts where medication names do not correspond to prescriptions (box 3, examples iii–v). In example iii, drug names are related to allergies: a specific module detects this relation and such drugs are not extracted. In

Box 2 Excerpts from four regular expressions for the extraction of mode (1, 2) and frequency (3,4) information

Pipe and parentheses allow disjunction of strings, while square brackets allow disjunction of characters, \n means end of lines and ? means an optional string or character, and back slash (\) is used to despecialize characters. Strings with the $ symbol indicate variables: they are described in the second part of this example. The first regular expression detects entities such as to each nostril, under the tongue, by mouth; the second expression detects nasal, drip, inhaled, subq; the third expression detects once a day, two per day, 2 per day; and the last expression detects bid, b.i.w.

1. ($adv)($sep)?($det|$adj)?($sep)?($anatomy)
2. (subcutaneously|subcutaneous|subcutane|subcu|subqutaneously|subqutaneous|subqtane|subq|inhaled|inh|iv|intravenously|intravenous|intraven|neb|drip|injection|inj|im|intramuscularly|intramuscular|intramusc)s?
3. ($number)($sep)?(a|per|$det)($sep)?d(ay)?
4. b($sep)?i($sep)?(d|w)

where the variables are:

adv = (through|per|by|with|via|in|to|under)
sep = ([\-\. \n])
det = (his|her|your|the)
adj = (each|both|right|left)
anatomy = (nose|eye|nostril|mouth)s?
number = ([0–9]+|once|one|two|twice|three|four)

Box 3 Examples of textual data to be processed

i. Heme. Anemia workup. Iron 49, TIBC 256, B12 555, folate normal, ferritin 102, reticulocyte 7.9, and Epogen level 19.
ii. HOME MEDS: methadone 20 bid, imdur 120 bid, hydral taking 25 bid, lasix 20 bid, coumadin, colace, iron, nexium 40 bid, doxazosin 2 qd, allopurinol 100 qod
iii. ALLERGY: prednisone, penicillins, tamsulosin, simvastatin, spironolactone
iv. … did not require medications for abdominal pain
v. INR's will be followed by Coumadin clinic; insulin-dependent diabetic; iron deficiency
vi. … Methadone 20 bid, Ofloxacin 200 mg p.o. q 12, Insulin lente 12 units subcu q p.m.
vii. … history of atrial flutter controlled on Amiodarone
viii. … started on calcitriol given elevation of parathyroid hormone
ix. … started on metoprolol 12.5 mg p.o. q.6 h. for rate control
x. … should be switched to Toprol as her blood pressure tolerates
xi. She was initially diuresed with IV Lasix.
xii. packed red blood cells, red blood transfusions, red blood cell, autologous red blood cells, blood, autologous blood, prbcs, prbc…
xiii. …, HCTZ 25 mg PO QD, Norvasc 10 mg PO QD, Pavachol 80 mg PO QD.

example iv, drugs occur in negative context, detected with a NegEx-inspired algorithm: it exploits the proximity of pre- and post-negation markers. In example v, drug names appear in other contexts: within names of diseases and institutions. This situation is processed through an extension of NegEx resources: proximity of terms such as 'clinic', 'dependent' or 'deficiency' allows these drugs to be not detected as prescriptions.

Detection of missing medication names

With the rapid evolution of therapeutic research, new drugs appear,[24] but drug nomenclatures cannot keep pace. We propose a novel method for a more exhaustive identification of new drugs. The main indication we rely on is that drugs often occur in specific semantic contexts together with medication-related information (box 3, example vi). The corresponding semantic pattern is: m do mo? f, where medication name, m ('methadone', 'ofloxacin', …), is followed by dosage, do ('20', '200 mg', '12 units'), possibly followed by administration mode, mo ('p.o.', 'subcu'), and followed by frequency, f ('twice daily', 'q 12', 'q p.m.'). If all entities (do, mo and f) except the first one are recognized, we infer that the first entity is a new drug name. We additionally check whether this entity is a stopword and whether its ending is typical of drug endings (eg, 'ine', 'one', 'ase', 'ate', 'cin', 'rin').

Identification of reasons

Reasons are identified by two approaches: (1) the use of terminological resources; (2) the use of noun phrase extraction together with reason markers. The first approach applies only Snomed International terms and patient complaints. The second approach allows the sensitivity of this vocabulary to be increased through extraction of noun phrases. However, exploiting all these noun phrases can be disastrous for precision, as the majority are not relevant for the reason category. Combining noun phrases with 52 contextual patterns ('for', 'given', 'controlled on', …) allows them to be constrained (examples vii–ix, box 3).

Evaluation

Evaluation was performed by organizers of the challenge: automatically generated results are compared with the 251 documents from ground truth according to the protocol described by Uzuner *et al*.[34] The main evaluation measure is the F-measure computed for exact and inexact matches.

RESULTS

Table 2 presents results for our system in terms of F-measure *F*, precision *P* and recall *R*. The global exact-match F-measure was 0.78. Within the challenge framework, our system ranked 7th out of 20 participating systems. The system generated good results (F-measure over 0.81) for four categories (drug, dosage, frequency, mode). The two remaining categories (duration and reasons) were extracted with lower performance (F-measure 0.36 and 0.29, respectively). Exact match performed slightly better than inexact match. Within the interval of medication occurrences,[2 11 6] the mean number of medications per document was 35.6. Only one document has no mention of drugs.

DISCUSSION

As shown in table 3, the performances obtained on the training (n=17) and test sets were comparable. Stability of the system was a positive result, especially given the very small set of annotated training data. We assume that the system may be useful for the processing of other clinical records, or at least can be easily adapted. Overall, it allows processing of narrative clinical documents and extraction of several medication-related

Table 2 Test set: performance of the system for exact and inexact matches

	Exact match			Inexact match		
	F	*P*	*R*	*F*	*P*	*R*
m	0.81	0.84	0.80	0.83	0.87	0.80
do	0.82	0.87	0.78	0.85	0.88	0.82
f	0.84	0.83	0.84	0.84	0.84	0.84
mo	0.87	0.85	0.88	0.86	0.84	0.88
du	0.36	0.35	0.37	0.36	0.37	0.35
r	0.29	0.30	0.27	0.34	0.44	0.28
Global	0.78	0.80	0.76	0.78	0.81	0.75

data with good performance, making the tedious manual annotation easier.

The core platform for NLP processing relies on standard NLP steps (NER, tokenization, part-of-speech (POS) tagging, lemmatization), but also on specific modules designed for this task. An original point—that is, tokenization performed after the NER—allows disambiguation of several cases where punctuation does not stand for sentence boundaries. Implementation of the tools and modules used within the Ogmios platform also facilitates communication between them, making the management of linguistic and semantic annotations easier.[39] In addition, the integration of modules with regular expressions is also easy and does not conflict with other modules and tools.

An analysis of these results was performed on 26 randomly selected discharge summaries from the ground truth (10%). Within this set, a total of 729 medication annotations were analyzed: 380 were identical and 47 overlapped with the reference annotations. In the remaining annotations, at least one category was different. This difference may correspond to false-positive (n=70) or false-negative (n=162) annotations.

We found only 16 (2%) false positives due to the extraction of wrong medication names, which attests to the quality of the drug lexicon. However, a few entries (ie, 'acute phase reactant', 'haemophilus influenzae', 'chewable') remained that were wrongly considered as drugs. The quality of medication lexica is a common problem[19 31]: with the original RxNorm, the F-measure falls to 40.73%. Early in our experience, we observed this fact and manually removed a large number of entries. Nevertheless, additional filtering is required. It cannot be done using a vocabulary of common English words, as in Sirohi and Peissig,[19] because nearly all these entries are relevant to the medical area: cleaning them up would instead require additional manual work or contextual rules. Another category of noise among the extracted drugs is related to ambiguous medication names that escaped our attention or for which the context is not indicative of their semantics. False positives within medication-related information are often due to wrong semantic relations.

Table 3 Training set: performance of the system for exact and inexact matches

	Exact match			Inexact match		
	F	*P*	*R*	*F*	*P*	*R*
m	0.86	0.85	0.87	0.85	0.83	0.86
do	0.80	0.83	0.78	0.82	0.84	0.80
f	0.83	0.80	0.87	0.82	0.81	0.83
mo	0.88	0.82	0.94	0.87	0.80	0.96
du	0.51	0.46	0.57	0.57	0.52	0.63
r	0.38	0.32	0.47	0.45	0.42	0.48
Global	0.79	0.76	0.82	0.78	0.76	0.80

The most commonly recurring problem is associated with reason detection: in examples x–xi (box 3), our system wrongly extracts 'blood pressure' as the reason for administration of 'toprol' and 'diuresed' as the reason for 'IV Lasix'.

We found several cases of false negatives among drug names:

1. Ambiguous drug names (eg, 'iron', 'statin', 'blood', 'magnesium', 'glucose') corresponding to administered products but not occurring in expected positive contexts
2. Terms such as 'fluids', 'agents' or 'medication' that we considered to be under-specified, but that should be extracted
3. Some classes of drugs (eg, 'antianginal therapy', 'pressure medications') missing from our resources
4. New drug names (eg, 'vp-16', 'ducolox', 'vasopressor', 'guqifenesin') that did not occur within expected semantic patterns
5. Misspellings and abbreviations (eg, 'aspirin325', 'hep.')
6. Pronominal phrases (eg, 'these medications')

Blood products remain difficult to detect, as they seldom appear within listings but mainly in narrative sections. Moreover, their nomenclature is not standardized, and various phrases are used to refer to a blood transfusion (box 3, example xii). An extension of semantic patterns may be helpful: 'required' and 'one unit of' are valuable indicators that 'blood' was administered in the phrase 'required one unit of blood during her hospital course'.

An additional analysis was performed of the module for detection of new medication names. It extracted 49 occurrences, 15 of which are real drug names (precision=30%), such as 'pendalol', 'lithium', 'permatol', 'levoxine' or 'pavachol' (box 3, example xiii). The precision is low, but it should be noted that we used it for enriching an already large drug nomenclature (over 240 000 entries) and it missed only a few occurrences (such as 'guqifenesin'). A more thorough evaluation of this module is ongoing.

Other false negatives correspond to missed drug-related information. It is seldom due to the incompleteness of the defined rules, but to wrong computation of dependency relations. Syntactic parsing[42 43] may be helpful for this.

CONCLUSION AND FUTURE WORK

We have described a system developed for the annotation and extraction of medication-related information from narrative discharge summaries. We looked at this task as an annotation and annotation-disambiguation problem. Specific semantic resources were exploited in a rule-based approach. We also proposed a novel module for detection of new medication names through the exploitation of semantic patterns. Global performances of our system (F-measure 0.78) rate it 7th among the 20 participants of the I2B2 challenge. Our system provides an F-measure of over 0.81 for extraction of medication names, their frequency, dosage and mode of administration; however, it performs poorly with duration and reasons, which is also the case for other participating systems.

Among the benefits are: improved duration extraction through exploitation of prepositional phrases; improved reason extraction with extended noun phrases; further evaluation of the module for deducing new medications; improved establishment of dependency relations between drug names and the related information.

Acknowledgments We are grateful to: the organizers of the I2B2 challenge for preparing and providing such an exciting framework for the evaluation of text mining systems; the anonymous reviewers for helpful and constructive comments; and Aurélie Névéol and Amandine Périnet for editorial assistance.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **WHO.** *International drug monitoring: the role of national centers*. Geneva, Switzerland: World Health Organization, 1972.
2. **FDA.** *Good pharmacovigilance practices and pharmacoepidemiologic assessement*. Rockville, MD: Food and Drug Administration, 2005.
3. **Pronovost P,** Weast B, Schwarz M, *et al*. Medication reconciliation: a practical tool to reduce the risk of medication errors. *J Crit Care* 2003;**18**:201—5.
4. **Bates D,** Leape L, Cullen D, *et al*. Effect of computerized physician order entry and a team intervention on prevention of serious medication errors. *JAMA* 1998;**280**:1311—16.
5. **Teich J,** Merchia P, Schmiz J, *et al*. Effects of computerized physician order entry on prescribing practices. *Arch Intern Med* 2000;**160**:2741—7.
6. **Oren E,** Shaffer E, Guglielmo B. Impact of emerging technologies on medication errors and adverse drug events. *Am J Health Syst Pharm* 2003;**60**:1447—58.
7. **Boussadi A,** Bousquet C, Sabatier B, *et al*. Specification of business rules for the development of hospital alarm system: application to the pharmaceutical validation. *Stud Health Technol Inform* 2008;**136**:145—50.
8. **Imming P,** Sinning C, Meyer A. Drugs, their targets and the nature and number of drug targets. *Nat Rev Drug Discov* 2006;**5**:821—34.
9. **Rader D,** Daugherty A. Translating molecular discoveries into new therapies for atherosclerosis. *Nature* 2008;**451**:904—13.
10. **Hale R.** Text mining: getting more value from literature resources. *Drug Discov Today* 2005;**10**:377—9.
11. **McDonald C,** Tierney W. The medical gopher—a microcomputer system to help find, organize and decide about patient data. *West J Med* 1986;**145**:823—9.
12. **Poon E,** Blumenfeld B, Hamann C, *et al*. Design and implementation of an application and associated services to support interdisciplinary medication reconciliation efforts at an integrated healthcare delivery network. *J Am Med Inform Assoc* 2006;**13**:581—92.
13. **Manley H,** Drayer D, McClaran M, *et al*. Drug record discrepancies in an outpatient electronic medical record: frequency, type, and potential impact on patient care at a hemodialysis center. *Pharmacotherapy* 2003;**23**:231—9.
14. **Grant R,** Devita N, Singer D, *et al*. Improving adherence and reducing medication discrepancies in patients with diabetes. *Ann Pharmacother* 2003;**37**:962—9.
15. **Thomsen L,** Winterstein A, Søndergaard B, *et al*. Systematic review of the incidence and characteristics of preventable adverse drug events in ambulatory care. *Ann Pharmacother* 2007;**14**:1411—26.
16. **Rindflesch T,** Tanabe L, Weinstein J, *et al*. EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac Symp Biocomput* 2008:517—28.
17. **Kolárik C,** Hofmann-Apitius M, Zimmermann M, *et al*. Identification of new drug classification terms in textual resources. *Bioinformatics* 2007;**23**:264—72.
18. **Chen E,** Hripcsak G, Xu H, *et al*. Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. *J Am Med Inform Assoc* 2008;**15**:87—98.
19. **Sirohi E,** Peissig P. Stufy of effect of drug lexicons on medication extraction from electronic medical records. *Pac Symp Biocomput* 2005:308—18.
20. **Levin M,** Krol M, Doshi A, *et al*. Extraction and mapping of drug names from free text to a standardized nomenclature. *AMIA Annu Symp Proc* 2007:438—2.
21. **Chhieng D,** Day T, Gordon G, *et al*. Use of natural language programming to extract medication from unstructured electronic medical records. *AMIA Annu Symp Proc* 2007:908—8.
22. **Cimino JJ,** Bright TJ, Li J. Medication reconciliation using natural language processing and controlled terminologies. *Stud Health Technol Inform* 2007:679—83.
23. **WHO.** *The use of stems in the selection of international nonproprietary names (inn) for pharmaceutical substances. Technical report*. Geneva: World Health Organization, 2006.
24. **Segura-Bedmar I,** Martinez P, Segura-Bedmar M. Drug name recognition and classification in biomedical texts. *Drug Safety* 2008;**13**:816—23.
25. **Xu R,** Morgan A, Das AK, *et al*. Investigation of unsupervised pattern learning techniques for bootstrap construction of a medical treatment lexicon. *Proceedings of the BioNLP 2009 Workshop* 63—70.
26. **Gold S,** Elhadad N, Zhu X, *et al*. Extracting structured medication event information from discharge summaries. *AMIA Annu Symp Proc* 2008:237—41.
27. **Turchin A,** Morin L, Semere L, *et al*. Comparative evaluation of accuracy of extraction of medication information from narrative physician notes by commercial and academic natural language processing software packages. *AMIA Annu Symp Proc* 2006:789—93.
28. **Jagannathan V,** Mullett C, Arbogast J, *et al*. Assessment of commercial nlp engines for medication information extraction from dictated clinical notes. *Int J Med Inform* 2009;**78**:284—91.
29. **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.
30. **Iglesias JE,** Rocks K, Jahanshad N, *et al*. Tracking medication information across medical records. *Proc AMIA Annu Fall Symp* 2009:266—70.
31. **Xu H,** Stenner S, Doan S, *et al*. MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc* 2010;**17**:19—24.
32. **Shah AD,** Martinez C. An algorithm to derive a numerical daily dose from unstructured text dosage instructions. *Pharmacoepidemiol Drug Saf* 2006;**15**:161—6.
33. **Hripcsak G,** Friedman C, Alderson P, *et al*. Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med* 1995;**122**:681—8.
34. **Uzuner Ö,** Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc* 2010;**17**:514—8.
35. **Simon L,** Wei M, Robin M, *et al*. Rxnorm: prescription for electronic drug information exchange. *IT Professional* 2005;**7**:17—23.
36. *RxNorm, a standardized nomenclature for clinical drugs. Technical report, National Library of Medicine*. Bethesda, Maryland, 2009. http://www.nlm.nih.gov/research/umls/rxnorm/docs/index.html.
37. **Côté RA,** Rothwell DJ, Palotay JL. *SNOMED Internationale—The systematized nomenclature of human and veterinary medicine. College of American Pathologists – American Veterinary Medical Association, Northfield*, 1993.
38. **Lussier YA,** Rothwell DJ, Côté RA. The SNOMED model: a knowledge source for the controlled terminology of the computerized patient record. *Methods Inf Med* 1998;**37**:161—4.
39. **Hamon T,** Nazarenko A, Poibeau T, *et al*. *A robust linguistic platform for efficient and domain specific web content analysis. In Proceedings of RIAO 2007* (electronic edition), Pittsburgh, USA. 2007.
40. **Aubin S,** Hamon T. *Improving term extraction with terminological resources*. In: Salakoski T, Ginter F, Pyysalo S, *et al*, eds. *Advances in natural language processing (5th International Conference on NLP, FinTAL 2006), number 4139 in LNAI*. Springer, August 2006:380—7.
41. **Tsuruoka Y,** Tateishi Y, Kim JD, *et al*. Developing a robust part-of-speech tagger for biomedical text. *LNCS* 2005;**3746**:382—92.
42. **Charniak E.** Immediate-head parsing for language models. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. 2001:124—31.
43. **Klein D,** Manning CD. *Accurate unlexicalized parsing. In: Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 2003:423—30.