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

Thierry Hamon,1 Natalia Grabar2,3

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, pharmacovigilance1 2 aims to prevent adverse drug effects. Medical3 4 and pharmacological5 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,6 7 while the scientific literature helps biologists carrying out research on new drugs.8 9 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,10 11 but it has been observed that this type of data is often incomplete or out of date12 13 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 literature16 17 or in clinical records.18 19 It has been observed that the quality of such nomenclatures must be controlled,20 as it has a direct impact on the quality of results. Approximate matching was proposed as a method of drug name recognition20 and shown to improve extraction results compared with dictionary-based exact matching. Other methods aim to identify new drug names through naming conventions21 22 or contextual rules.23 Previous work has also addressed the extraction of drug-related information. The first study of this kind24 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 research25 26 27 and commercial28 29 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 entries30 significantly improves results (up to F-measure of 0.98). Common difficulties are related to incompleteness of drug lexica20 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.

1UM6BIO (EA3969), UFR SMBH Léonard de Vinci, Université Paris 13, 93017 Bobigny Cedex, France 2Centre de Recherche des Cordeliers, Université Pierre et Marie Curie - Paris6, UMR 7872-20, Paris, F-75006; Université Paris Descartes, UMR 7872-20, Paris, F-75006; INSERM, U872-20, Paris F-75006, France 3HEGP 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

Downloaded from https://academic.oup.com/jamia/article-abstract/17/5/549/831598 by guest on 08 March 2019
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. 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 290245 entries): drug nomenclature and pathology terms. We created a medication list containing 248 869 entries mainly provided by RxNorm. 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 proved to be an efficient and user-friendly source for NLP processing: 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.

### 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’, ‘cann’t’, ‘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 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, ‘upper case 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 was also 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: ‘qhs’ 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.

### 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)

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

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 biological characteristics or drugs. They were first assigned a specific semantic tag. Then, if they occurred in listings or medication-related sections (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
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 ‘clinical’, ‘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 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 together with reason markers. The following 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. 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 0.81) for four categories (drug, dosage, frequency, mode). The two remaining categories (duration and reasons) were extracted with lower performance (F-measure 0.56 and 0.29, respectively). Exact match performed slightly better than inexact match. Within the interval of medication occurrences, 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
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 40.75%. 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.

The most commonly recurring problem is associated with reason detection: in examples x–xi (box 5), our system wrongly extracts ‘blood pressure’ as the reason for administration of ‘toprol’ and ‘diuresed’ as the reason for ‘1V Lasix’.

We found several cases of false negatives among drug names:

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.
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 5, 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=50%), such as ‘pendalol’, ‘lithium’, ‘permatal’, ‘levoxine’ or ‘pavachol’ (box 5, 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.

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

<table>
<thead>
<tr>
<th></th>
<th>Exact match</th>
<th>Inexact match</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>P</td>
</tr>
<tr>
<td>m</td>
<td>0.81</td>
<td>0.84</td>
</tr>
<tr>
<td>do</td>
<td>0.82</td>
<td>0.87</td>
</tr>
<tr>
<td>f</td>
<td>0.84</td>
<td>0.83</td>
</tr>
<tr>
<td>mo</td>
<td>0.87</td>
<td>0.85</td>
</tr>
<tr>
<td>du</td>
<td>0.36</td>
<td>0.35</td>
</tr>
<tr>
<td>r</td>
<td>0.29</td>
<td>0.30</td>
</tr>
<tr>
<td>Global</td>
<td>0.76</td>
<td>0.80</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Exact match</th>
<th>Inexact match</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>P</td>
</tr>
<tr>
<td>m</td>
<td>0.86</td>
<td>0.85</td>
</tr>
<tr>
<td>do</td>
<td>0.80</td>
<td>0.83</td>
</tr>
<tr>
<td>f</td>
<td>0.83</td>
<td>0.80</td>
</tr>
<tr>
<td>mo</td>
<td>0.88</td>
<td>0.82</td>
</tr>
<tr>
<td>du</td>
<td>0.51</td>
<td>0.46</td>
</tr>
<tr>
<td>r</td>
<td>0.38</td>
<td>0.32</td>
</tr>
<tr>
<td>Global</td>
<td>0.79</td>
<td>0.76</td>
</tr>
</tbody>
</table>

**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 Néveu and Amandine Périnet for editorial assistance.

**Competing interests** None.
Provenance and peer review  Not commissioned; externally peer reviewed.

REFERENCES