Medication information extraction with linguistic pattern matching and semantic rules

Irena Spasić,[1] Farzaneh Sarafraz,[2] John A Keane,[2] Goran Nenadić[2]

▸ An additional appendix is published online only. To view these files please visit the journal online (http://jamia.bmj.com).

[1]Cardiff School of Computer Science & Informatics, Cardiff University, Cardiff, UK
[2]School of Computer Science, University of Manchester, Manchester, UK

Correspondence to
Dr Irena Spasić, Cardiff School of Computer Science & Informatics, Cardiff University, 5 The Parade, Roath, Cardiff CF24 3AA, UK;
i.spasic@cs.cardiff.ac.uk

Received 4 February 2010
Accepted 25 June 2010

ABSTRACT

Objective This study presents a system developed for the 2009 i2b2 Challenge in Natural Language Processing for Clinical Data, whose aim was to automatically extract certain information about medications used by a patient from his/her medical report. The aim was to extract the following information for each medication: name, dosage, mode/route, frequency, duration and reason.
Design The system implements a rule-based methodology, which exploits typical morphological, lexical, syntactic and semantic features of the targeted information. These features were acquired from the training dataset and public resources such as the UMLS and relevant web pages. Information extracted by pattern matching was combined together using context-sensitive heuristic rules.
Measurements The system was applied to a set of 547 previously unseen discharge summaries, and the extracted information was evaluated against a manually prepared gold standard consisting of 251 documents. The overall ranking of the participating teams was obtained using the micro-averaged F-measure as the primary evaluation metric.
Results The implemented method achieved the micro-averaged F-measure of 81% (with 86% precision and 77% recall), which ranked this system third in the challenge. The significance tests revealed the system's performance to be not significantly different from that of the second ranked system. Relative to other systems, this system achieved the best F-measure for the extraction of duration (53%) and reason (46%).
Conclusion Based on the F-measure, the performance achieved (81%) was in line with the initial agreement between human annotators (82%), indicating that such a system may greatly facilitate the process of extracting relevant information from medical records by providing a solid basis for a manual review process.

The 2009 i2b2 medication extraction challenge[1] focused on the extraction of medication-related information including: medication name (m), dosage (do), mode (mo), frequency (f), duration (du) and reason (r) from discharge summaries. In other words, free-text medical records needed to be converted into a structured form by filling a template (a data structure with the predefined slots)[2] with the relevant information extracted (slot fillers). For example, the following sentence:

"In the past two months, she had been taking Ativan of 3–4 mg q.d. for anxiety."

should be converted automatically into a structured form as follows:

m="ativan" || do="3–4 mg" || mo="nm" || f="q. d." || du="two months" || r="for anxiety"

Note that only explicitly mentioned information was to be extracted with no attempt to map it to standardized terminology or to interpret it semantically.

METHODS

Documents are processed in three steps (see figure 1): linguistic preprocessing, dictionary and pattern matching, and template filling.

Linguistic preprocessing

The goal of linguistic preprocessing is to annotate a document with relevant lexicosyntactic information. It includes sentence splitting,[3] part-of-speech (POS) tagging[4] and shallow parsing.[5] The output produced is an XML document in which XML tags are used to mark up sentences, POS categories of individual tokens (eg, nouns, verbs) and syntactic categories of word chunks (eg, noun phrases, verb phrases). We used the Penn Treebank tag set[6] for the annotations used throughout this article (eg, NN, NP, PP).

Dictionary and pattern matching

Documents are analyzed with rules that exploit morphologic, lexical and syntactic properties of the targeted information. As medication names are pivotal for the given information extraction task, we describe their recognition in more detail. We also focus on the recognition of reason as it proved most difficult to model.

Our approach to medication name recognition is primarily dictionary based. We assembled a dictionary of medication names semi-automatically. Medical records were split into sections by matching the most frequent title keywords, based on whether the sections were classified as: diagnosis, medical history, social/family history, physical/laboratory examination, medication list or other.[7] The content of medication list sections was analyzed, medication names extracted and added to a dictionary. Additional names were collected from relevant web pages. We shared the lexical resources assembled via the organizers.[1] We also exploited the morphology of medication names. We compiled a list of typical affixes (eg, -cycline, -nazole, sulfa-), which were matched against individual tokens, for example matching the suffix -statin extracted atorvastatin, lovastatin, nystatin, etc. Medication names were further expanded to include their form, release, strength, etc. The following are examples of fully expanded medication names: atrovent inhaler, metoprolol succinate extended release, procardia XL, nitroglycerin 1/150, maalox plus extra strength, etc.

Two other dictionaries were assembled to recognize generic medication types: (1) nouns that refer

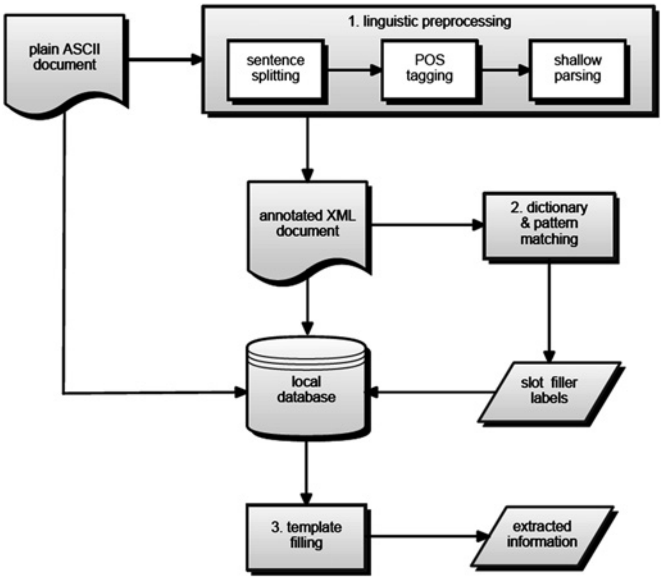

Figure 1 The system architecture diagram.

to medications in general (eg, medication, medicine, drug) and (2) modifiers that imply medication type (eg, antidepressant, blood pressure, cardiovascular). Medication types were acquired from the training dataset by automatically searching for NPs with a head noun (the main noun that is modified by other elements in an NP) from the dictionary (1) and manually curating the modifiers extracted. We processed the UMLS[8] terms with a head noun from the dictionary (1) in the same way. The two dictionaries were combined to recognize different classes of medications, for example, anti-inflammatory drugs, pain medicines, antihypertensive medications, cardiac meds, etc. A controlled list of modifiers prevented the extraction of false positives with similar lexicosyntactic properties, for example, admission medications, home meds, new medicines, recreational drugs, etc.

Recognition of the reason for medication followed an approach that explored nouns and adjectives frequently used to specify diseases and procedures.[9] We compiled the dictionaries of reason-related words based on a subset of UMLS terms from the relevant semantic types: disease or syndrome, therapeutic or preventive procedure, etc. We selected approximately 150 nouns and/or affixes (eg, discomfort, deficiency, swelling, cough, hypo-, hyper-, -pathy, -itis) and approximately 70 adjectives (eg, chronic, acute, abnormal, atypical, mild, severe, irregular) that can be used to identify a reason. Furthermore, an equivalent of the Latin phrase pro re nata (ie, PRN, P.R.N., as needed) is often followed by the reason, for example, aspirin 600 mg p.o. q4h PRN pain. We automatically extracted the right contexts of such expressions and used them to compile a dictionary of reason phrases as they occurred in the training dataset (eg, upset stomach, SOB, shortness of breath). Finally, a reason is identified either as an exact match on a reason phrase or as an NP that contains a reason-related noun or adjective.

In addition to morphological and lexical characterization of the reason candidates, we exploited their syntactic relation to the co-occurring medication. Medication is often followed by the reason specified as a PP with the preposition for, for example '… the patient was given cefuroxime and levofloxacin in the emergency department <PP>for a presumed community acquired pneumonia</PP>…'. Using this rule, we extracted examples such as trace edema, other etiology of her pneumonia, empiric treatment of urinary tract infection, deep vein thrombosis prophylaxis, etc.

We applied a similar strategy to assemble relevant dictionaries and define patterns for other types of information. All rules were implemented as expressions in Mixup (My Information eXtraction and Understanding Package),[10] a simple pattern-matching language. Matching Mixup expressions against text spans (ie, token sequences) produces a set of labels, a type of stand-off annotation that associates a span and its category (ie, the rule it matches). We used labels to encode POS tags (eg, NN, CD) or chunk parses (eg, NP) acquired during linguistic preprocessing. These labels were exploited by Mixup rules (eg, a cardinal number followed by a noun from a dictionary of units is labeled as a dosage), which produced an additional set of semantic labels (eg, do for dosage), illustrated here by XML tags: <do><NP><CD>325</CD> <NN>mg</NN></NP></do>

Template filling

Once the labels for potential slot fillers are produced, they are combined to fill the information extraction template with the following slots: medication, dosage, mode, frequency, duration and reason. Hereafter, we use the following XML tags to mark up the slot fillers: <m>, <do>, <mo>, <f>, <du> and <r>. A template is filled in three steps.

Step 1: Label filtering. We start by filtering the labels in order to merge consecutive labels of the same type, delete nested labels, and delete all negated medications and medications found in the allergy context or a laboratory examination section.

Step 2: Filling the medication slot. We start filling the template by importing all remaining labeled medications. They are further filtered using their context. For instance, if the word iron co-occurs with any of the words deficiency, deficient, insufficiency, insufficient, test or study, then that particular mention is not considered to be a prescribed medication, for example, in the following sentence: 'The patient was found to be <m>iron</m> deficient and she was continued on <m>iron supplements</m> three times a day.', the first mention of iron is removed from the template.

Step 3: Filling other slots. We look at the medication's immediate context between the nearest co-occurring medications within the given sentence. We use heuristic rules that use clues such as proximity and punctuation to associate the medication with the most appropriate labels (ie, dosage, mode, frequency and duration) from the given context.

Filling the reason slot

Associating a medication with a correct reason is slightly more complex, as their syntactic relationship is highly variable and often indirect, when the reason may not be even specified in the same sentence, for example:

'The patient had evidence of <r>pneumonia</r> on his chest x-ray. He was started on <m>antibiotics</m>.'

Therefore, in cases in which there is no link between a medication and a reason by means of PRN or a PP (see section on Dictionary and pattern matching), we analyzed the whole sentence as well as the preceding one, and applied a set of rules to associate them.

Semantic rules

We divided medications into 18 types: diuretics, anticoagulants, blood pressure medications, etc. Each medication type (eg, diuretics) given as a list of medication names (eg, lasix, furosemide, aldactone) was associated with typical indications given as a list of expressions describing them (eg, diuresis, swelling, renal insufficiency). A medication is mapped to its type, after which the associated indications are matched against the medication's context. If a match is found, then an NP that contains a given expression is extracted as the reason. For example, in the following sentence:

'She was restarted on <m>Aldactone</m> with <NP>effective gentle diuresis</NP>.' Aldactone was recognized as a diuretic, after which the word diuresis from the list of indications was found, and the reason expanded to the NP that contains it, that is, effective gentle diuresis.

Single medication rule

We look for sentences with a single medication assuming that if a potential reason is found it is likely to correspond to the given medication. If no reason is found within the same sentence, we consider the predecessor sentence. If it contains a medication name, then we discard the sentence assuming that any reason mentioned is more likely to refer to the medication in that particular sentence rather than its successor. Otherwise, all reasons found are associated with the medication, for example:

'Hematocrit checked at that time revealed <r>a significant drop in blood count</r> to 26. The patient was transfused two units of <m>packed red blood cells</m>.'

Anaphora resolution

Phrases such as the pain, this pain, her pain and his pain were often extracted as reasons. While in most cases they do represent a reason, they typically refer to more detailed preceding utterances. In such cases, we attempt to resolve the anaphora (ie, references to previous entities in the discourse). Whenever the/this/her/his pain is extracted, we look for a previously occurring NP that contains the word pain. If found, it is used to replace the reason. In the following example, nitroglycerin was originally associated with the pain:

'The patient had <NP>recurrent chest pain</NP> which radiated to her right jaw. <r>The pain</r> was relieved with sublingual <m>nitroglycerin</m>.'

It is then replaced by the more informative NP recurrent chest pain found in the preceding sentence.

Deleting side effects

Side effects lexically and syntactically coincide with indications, which occasionally results in incorrectly labeling them as reasons. We applied a set of semantic rules to map medications to possible side effects, which can then be used to filter out some of the false positives suggested as reasons. For example, all NPs that contain a word with a prefix hypo- are labeled as potential reasons. Therefore, using the single medication rule, aspart would be incorrectly associated with the symptoms of hypoglycemia as a reason in the following sentence:

'His next dose of <m>aspart</m> should be either halved or held as the patient easily becomes hypoglycemic and is unable to recognize <r>the symptoms of hypoglycemia</r> himself.'

However, knowing that hypoglycemia is a likely side-effect of aspart as well as other antidiabetic agents (eg, lantus, humulin, novolin) is used to remove the false positive.

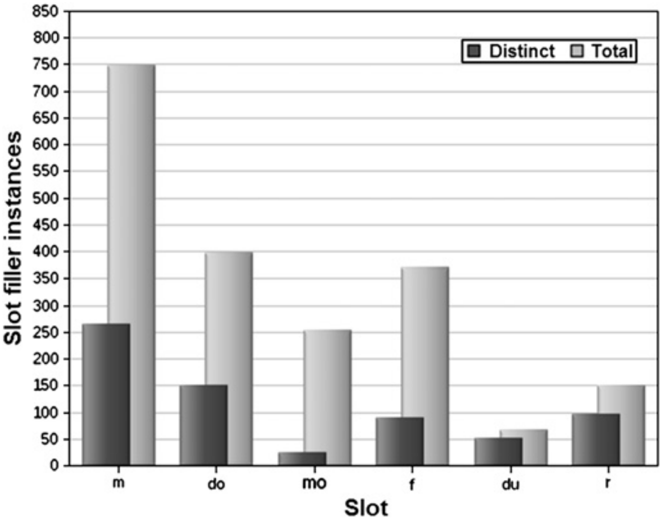

Figure 2 Distribution of annotations in the gold standard: total and distinct number of slot filler instances.

EXPERIMENTS AND RESULTS

Experimental setting

A total of 696 documents was distributed during the development phase (training data), followed by an additional set of 547 documents released for the evaluation purposes (test data). A subset of 17 training documents was annotated manually by the organizers. The final evaluation was performed on the gold standard of 251 test documents annotated manually by the participants. Figure 2 provides the distribution of manual annotations in the gold standard. The performance was evaluated using recall (R) and precision (P), as well as their combination into the F-measure.[1] These values were micro-averaged across each slot (vertical evaluation) as well as the whole entries that take into account the links between the slot fillers (horizontal evaluation). Micro-averaged horizontal evaluation was used as the primary evaluation metric to rank the results.

Results

We submitted the results of three system runs (table 1), which differed as to which reason extraction rules were applied: run 1 used all rules described in the section on filling the reason slot; run 2 used all rules from run 1 apart from the single medication rule; run 3 used all rules from run 2 apart from the semantic rules.

The best results were achieved in run 2 (F-measure of 81%), based on which our system was ranked third. This was not significantly different from the result of the second ranked system (F-measure of 82%).[1] These results were in line with the initial agreement between human annotators, whose pairwise micro-averaged F-measure was 82% on the gold standard annotated by two independent annotators. Relative to other systems, our

Table 1 Evaluation results for the three submitted runs

	Run 1			Run 2			Run 3		
Slot	P	R	F	P	R	F	P	R	F
m	88.24	79.78	83.80	88.38	79.78	83.86	88.44	79.76	83.87
do	90.60	78.27	83.99	90.68	78.29	84.03	90.72	78.29	84.05
mo	87.76	81.59	84.56	87.82	81.59	84.59	87.84	81.56	84.58
f	88.17	82.75	85.37	88.26	82.72	85.40	88.30	82.70	85.41
du	59.81	46.55	52.35	60.24	46.55	52.51	60.52	46.55	52.62
r	51.63	41.52	46.03	56.64	38.48	45.82	60.64	29.88	40.03
X	85.66	76.75	80.96	86.38	76.53	81.16	87.05	75.90	81.09

X in the last row indicates a complete entry in the template, that is horizontal evaluation.

system achieved the highest F-measure for the extraction of duration (53%) and reason (46%). Due to the strict annotation guidelines many semantically correct reason examples were counted as errors. Conforming to the guidelines improved both precision (52% to 56%) and recall (42% to 44%) of reason extraction, and thus the F-measure (46% to 49%). Even though the best reason extraction results were achieved in run 1, the best overall performance was achieved in run 2. This implies that the single medication rule may increase the number of correctly recognized reason candidates, but does not necessarily associate them with the correct medication, while semantic rules provide a good basis for associating medications and their reasons.

Figure 3 compares the performance against the training data (17 documents) and the test data (251 documents). Naturally better performance against the training data can be attributed partly to the medication name recognition module, as the training dataset was used as the main source to compile a dictionary. A failure to recognize a medication name (at least partially) would automatically result in a failure to extract the associated information. This consequently decreases the F-measure achieved for each slot individually as well as the overall F-measure. Figure 4 illustrates this point by comparing the F-measure achieved on the test data by projecting the results onto the recognized medication names only, in which we see consistent improvement across all slots when the corresponding medication name is extracted.

CONCLUSIONS

Machine learning approaches to information extraction may offer easier portability solutions, while rule-based information extraction systems tend to provide reliable results.[2] However, they tend to rely on hand-crafted rules tailored for a specific domain and a task at hand, and thus are not readily re-usable. Nonetheless, most machine learning approaches are probabilistic in their nature, and as such require significant amounts of training data, which are typically expensive to obtain. For example, the first ranked system was machine learning based and trained on a subset of 145 manually annotated documents. On such a scale, the annotation process may be as laborious as the rule engineering. Given that the annotations were provided for 17 training documents only, we opted for a rule-based approach. We successfully demonstrated the potential of rapidly developing high-performing rule-based information extraction systems through semi-automatic analysis of the training data and re-use of domain knowledge from public resources. The performance of our rule-based system was not only comparable to that of the top ranked systems, but also proved superior in extracting relatively sparse information such as duration and reason. Given the short development timescale, we strongly believe that significant improvements are possible, and that performance with an F-measure well over 80% if not 90% seems realistic. The optimal compromise between performance and portability should be achieved by an appropriate combination of machine learning and rule-based approaches.

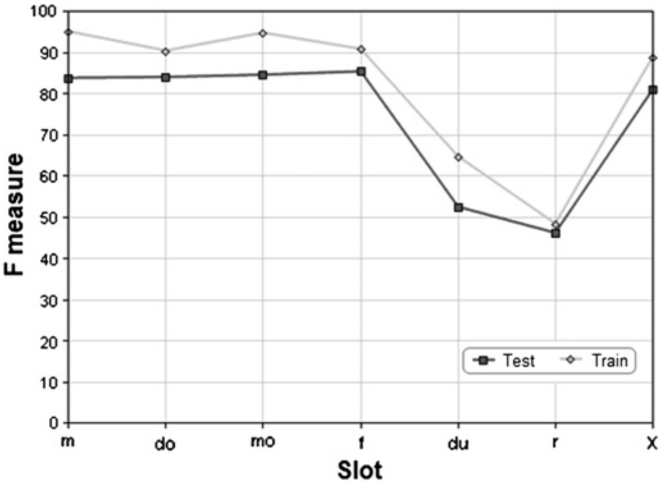

Figure 3 Comparison of the results achieved on the training and test data.

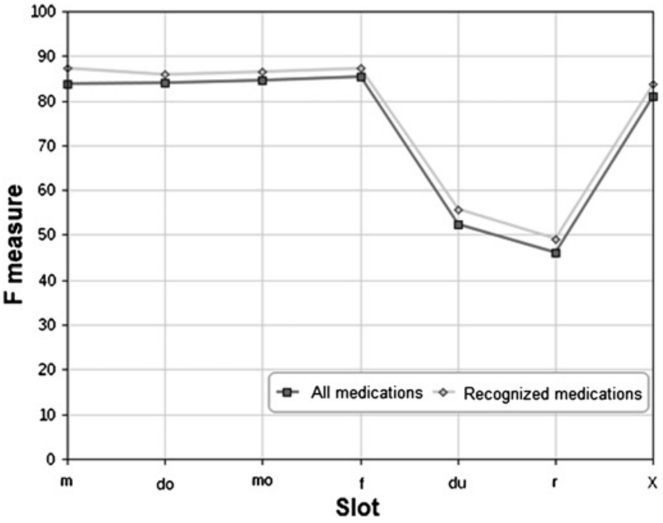

Figure 4 Overall improvement of the results for correctly recognized medication names.

Funding IS was supported by the BBSRC/EPSRC grant 'the Manchester Centre for Integrative Systems Biology'. GN and JAK were partially supported by the BBSRC grant 'Mining Term Associations from Literature to Support Knowledge Discovery in Biology'. 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. **Uzuner Ö,** Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assn* 2010;**17**:514—18.
2. **Cowie J,** Lehnert W. Information extraction. *Commun ACM* 1996;**39**:80—91.
3. A highly accurate sentence and paragraph breaker. http://text0.mib.man.ac.uk:8080/scottpiao/sent_detector (accessed 22 Jan 2010).
4. **Tsuruoka Y,** Tsujii J. Bidirectional inference with the easiest-first strategy for tagging sequence data. In: *Proceedings of the Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Vancouver, Canada, 6—8 October 2005:467—74.
5. **Tsuruoka Y,** Tsujii J. Chunk parsing revisited. In: *Proceedings of the 9th International Workshop on Parsing Technologies*. Vancouver, Canada, 9—10 October 2005:133—40.
6. **Marcus MP,** Marcinkiewicz MA, Santorini B. Building a large annotated corpus of English: the Penn Treebank. *Comput Linguist* 1993;**19**:313—30.
7. **Yang H,** Spasić I, Keane J, *et al*. A text mining approach to the prediction of a disease status from clinical discharge summaries. *J Am Med Inform Assn* 2009;**16**:596—600.
8. **Bodenreider O.** The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**:D267—70.
9. **Bodenreider O,** Burgun A, Rindflesch TC. Assessing the consistency of a biomedical terminology through lexical knowledge. *Int J Med Inform* 2002;**67**:85—95.
10. **Cohen WW.** MinorThird: methods for identifying names and ontological relations in text using heuristics for inducing regularities from data. http://www.minorthird.sourceforge.net (accessed 22 Jan 2010).