Automatic extraction of medication information from medical discharge summaries

Hui Yang

▶ Additional data are published online only. To view these files please visit the journal online (http://jamia.bmj.com).

Department of Computing, Open University, Milton Keynes, UK

Correspondence to
Dr Hui Yang, Department of Computing, Faculty of Mathematics Computing and Technology, The Open University, Walton Hall, Milton Keynes MK7 6AA, UK; h.yang@open.ac.uk

Received 20 February 2010
Accepted 30 June 2010

ABSTRACT

Objective This article describes a system developed for the 2009 i2b2 Medication Extraction Challenge. The purpose of this challenge is to extract medication information from hospital discharge summaries.
Design The system explored several linguistic natural language processing techniques (eg, term-based and token-based rule matching) to identify medication-related information in the narrative text. A number of lexical resources was constructed to profile lexical or morphological features for different categories of medication constituents.
Measurements Performance was evaluated in terms of the micro-averaged F-measure at the horizontal system level.
Results The automated system performed well, and achieved an F-micro of 80% for the term-level results and 81% for the token-level results, placing it sixth in exact matches and fourth in inexact matches in the i2b2 competition.
Conclusion The overall results show that this relatively simple rule-based approach is capable of tackling multiple entity identification tasks such as medication extraction under situations in which few training documents are annotated for machine learning approaches, and the entity information can be characterized with a set of feature tokens.

INTRODUCTION

The 2009 i2b2 medication extraction challenge[1] was organized to evaluate natural language processing (NLP) systems for the extraction of medication-related information from hospital patient reports. This challenge focused on the identification of medications and their associated attributes such as dosage, frequency, mode/route of administration, duration, reason for administration and list/narrative, which are called 'medication constituents' in our paper. A total of 696 discharge summaries was released during the development period. Only a tiny set of 17 hand-annotated samples by the i2b2 organizers was provided together with an annotation guideline[1] that was used by the participant teams to help curate their own annotation set. Another set of 547 discharge summaries was held out for the testing, and the system evaluation was performed on the subset of 251 clinical documents that were manually annotated by the participant teams after the competition.

This article describes our effort in the 2009 i2b2 medication extraction task. We investigated the effectiveness of a relatively simple rule-based approach to the extraction task. This approach incorporated several NLP techniques including term-based identification, context-related pattern detection and token-based lexical resource construction.

METHODS

Our system approached the medication extraction task by a three-step process, text pre-processing, medication constituent recognition and medication information integration, respectively. The basic processing procedure is depicted in Figure 1, which was coupled with several lexical resources that captured orthographic or morphological variations within medication constituents, and function components such as regular expressions, context-related patterns and constituent-related token classes.

Text preprocessing

- Section categories: the summary text was chunked into six predefined section categories. Section categories were recognized by the surface features of the headings (eg, 'diagnosis', 'examination', 'medication', etc.).
- Basic textual information: the text was decomposed into text lines, and tokenization was performed and relevant token position information was stored.
- Syntactic information: part-of-speech tagging and phrase recognition were performed by the GeniaTagger tool.[2]

Construction of lexical resources

Due to the lack of annotated training documents, all of the hand-crafted lexical resources used for this challenge were directly built from the training data. We did not use any external medical resources or knowledge bases such as UMLS[3] and SNOMED-CT.[4] We only chose the training data for the construction of lexical resources because it required limited development effort and was much more straightforward. We were interested in seeing how effectively it would perform on the testing data without relying on any external medicine resources.

Each of the collected lexicon resources included a set of feature terms in terms of one medication constituent category. For feature term selection, we used a combination of guided and manual methods. Most features were identified by medical relevance, and were linked to some specific medication constituent category. A more detailed discussion about lexical resource construction is available in a *JAMIA* data supplement, available online only.

1. Medication name lexicon. The lexicon included medications with brand names, generic names and common abbreviations. Some special terms,

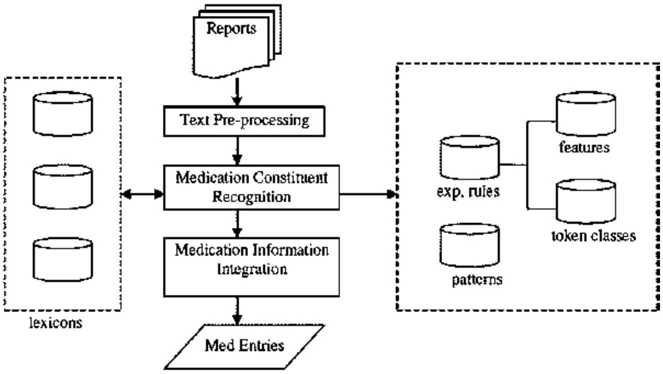

Figure 1 The system framework for medication extraction

such as 'regular' (refer to 'regular insulin'), 'anesthetic medication' (a drug class) and 'this medication' (co-reference), were also manually collected from the training data based on the annotation guideline. Once manual collection was completed, we developed a regular-expression-driven string replacement approach to cope with morphologic variations between mediation names in order to further enrich the completeness of the medication list:

- Full name versus abbreviation (eg, 'asa' for 'aspirin')
- Synonyms (eg, 'nicotinic acid sustained release', 'nicotinic acid sust. rel.' and 'nicotinic acid sr')
- Single medication with two names and one of the names is parenthetical (eg, 'advil (ibuprofen)' → 'advil' and 'ibuprofen')
- Spelling variants (eg, 'dovonex cream' and 'dovonex')

2. Dosage lexicon. Dosage information indicates the amount and unit in a medication administration. It generally appears in the form of a numeric with a noun word, for example, 'one tablet', '0.5 mg'. Table 1 shows a list of the morphological features used in the recognition of numeral dosage representation. A number of regular expression rules that geared towards the orthographic features of numeral dosage representation were applied by combining with 95 amount/unit nouns (eg, 'tab', 'mg', etc.). In addition, we also collected a list of non-numeric forms of dosage mentions such as 'small dose', 'sliding scale', etc.
3. Frequency lexicon. We observed that quite a number of frequency mentions (eg, 'q.3 h', '2×week', etc.) could be lexically split into different token classes. Table 2 depicts a set of frequency token classes that characterize the features of frequency mentions. A number of token-based rules presented in Table 3 was applied to build a set of morphological forms used for the extraction of frequency mentions. Moreover, a small set of special terms (eg, 'qmwf', 't/wed/th/sat/sun'), which did not conform to the token-based rules, were collected from the training data to supplement the token-based rule matching approach.
4. Mode/route lexicon. One hundred and eighteen mode terms that describe the method for administering the medications were selected from the training data. The term list included acronyms (eg, 'sl' for 'sublingual') and morphologically variant terms (eg, 'p o', 'p.o.' and 'po').
5. Duration lexicon. Our system classified the duration mentions into two types of expressions based on the analysis of the annotated sample reports and the annotation guideline:
   - Numeral-related expressions: such expressions always reveal the dosage amount information or administration length, for example, '30 pack' and '×10 days'. To extract such types of duration expressions, we adopted a similar token-based approach described in frequency extraction (also see Tables 2 and 3).
   - Non-numeral expressions: due to the complexity and irregularity of non-numeral expressions with PP forms (eg, 'through May', 'while at rehab only'), it is hard to find relevant morphological patterns to illustrate them. Therefore, we had to manually gather a small set of non-numeral expressions from the training data.
6. Reason lexicon. Reason mentions usually link to a variety of diseases and their associated symptoms and treatments. It means that a huge medical vocabulary dictionary is needed so as to cover the complexity and variations in reason expressions. Due to the development time limit and the sparseness of the annotation data, reason expressions were simply selected from the training data by hand.

Table 1 Morphological features for numeral representation in dosage terms

Feature	Example	Feature	Example
Digit	20	DigitSlashDigit	2/3
DigitAlpha	20 mg	DigitSlashDigitAlpha	500/50 mg
DigitAlphaDigitAlpha	7.5 mg×3d	DigitAlphaSlashAlpha	0.4 mg/spray
DigitHyphenDigit	2 to 3	Cardinal	One
DigitHyphenAlpha	2-litre	Cardinal to Cardinal	One to two
Digit Comma Digit	1000	CardinalHyphenOrdial	One-third
Digit to Digit	2 to 3	Word_num	Several

Table 2 Frequency token classes and their token elements

Token class	Token element
Meal	breakfast, lunch, dinner, supper, meal, meals
Daytime	a.m., am, p.m., pm, morning, afternoon, noon, evening, night, nighttime, bedtime
Weekday	Monday, Mon, Tuesday, Tues, Wednesday, Wed, Thursday, Thurs, Friday, Fri, Saturday, Sat, Sunday, Sun
TimeUnit	hour, h, hr, minute, min, day, d, week, wk, month
OneTimeUnit	a day, one day, per day, daily, a week, one week, per week, weekly, a month, one month, monthly
Num	3, 3/5, 1–2, 6 to 8, one, one to three,
Latin	qd, q.d., bid, b.i.d., tid, t.i.d., qid, q.i.d., qod, q.o.d, biw, b.i.w, hs, h.s, qhs, ac, a.c
AsNeed	as necessary, as needed, as directed, prn, p.r.n.

Table 3 Part of token-based morphological rules for frequency expression

Token-based rule	Example
[after|before|at|following|with|w/]+<Meal>	after breakfast, before meals, at supper, following lunch
[in|on|at|during]+<Daytime>	in the a.m., at bedtime, on p.m., during the evening
[each|every|on]+<Weekday>	each Monday, every Sunday, on tues,
[every]+<Num>+<TimeUnit>	every 3 hour, every 3–5 min
<Num>+[x|x/]+<TimeUnit>	2×/wk, 2–3×/day, 2×wk
[q|q.]+<TimeUnit>	qhr, q day, q.wk, q. week
[q|q.]+<Num>+<TimeUnit>	q2h, q 4 h, q. 2 weeks, q.6 h
[q|q.]+<Meal>	qlunch, q breakfast, q.meal, q. dinner
[q|q.]+<Daytime>	qam, q p.m., q. afternoon, q.evening
[q|q.]+<Weekday>	qwed, q monday, q. friday, q.saturday
[once|twice] + <OneTimeUnit>	once a day, twice per day
<Num>+[times|x] +<OneTimeUnit>	2 times a day, 3×daily

Extraction of medication constituent information

1. Term-based identification

   Most of medication information (ie, medication name, dosage, mode, frequency, duration and reason) was extracted using the term-based and token-based rule matching approach. In the term identification step, the feature terms stored in the lexical resources were converted to regular expressions for matching against the discharge summaries. The regular expressions were generally case-insensitive and matched whole words. Once a match was found in text, it was marked with the corresponding medication constituent tag (eg, 'm', 'do', 'mo', 'feq', 'du', etc.).
2. Post-processing for medication mention identification

   After the term identification step, we employed a context-based filter-out step to exclude the medication mentions that occurred in the following contexts:
   - Allergy contexts (eg, 'allg: sulfa/amoxicillin')
   - Negation contexts[5] (eg, '… was not on plavix')
   - Family members (eg, 'the patient' father was on Coumadin for ten years')
   - Homonymous to non-drug terms (eg, 'coumadin' in 'Coumadin clinic').

     In addition, some special medications, such as blood components for transfusion (eg, 'red blood cells (rbc)') and ingredients for total parenteral nutrition (eg, 'magnesium' and 'sodium'), were considered as valid candidates only when they were mentioned in the following contexts:
   - Medication list (eg, 'continued asa, bb, ace, rbc, wbc, …')
   - Collocated with other medication constituents such as dosage, frequency, etc. (eg, 'he was given 20 mg of sodium')
   - Presence pattern contexts (eg, '… continue on magnesium …')
3. Extraction of list/narrative

   Our approach to determining whether a medication appears in the medication list text was made up of two steps:
   - Section-level: examine the section category type of the text in which the medication occurs, which should be tagged as 'medication/disposition' in the text preprocessing step mentioned earlier.
   - Sentence-level: search the keyword clues, such as 'medications included …' and 'meds: …', in the text preceding the medication mention.

Integration of medication information

When a medication mention was found, the neighborhood around the medication was examined to look for the presence of other relevant medication constituents, for example, dosage, frequency, mode, duration and reason. The neighborhood was bounded by a ±2 line text window according to the annotation guideline. However, to improve the accuracy of integration, we added several additional constraints to further refine the search:

1. The order for the search should be started from the medication mention position. First forward search for the right context, and then backward search for the left context.
2. If there is another medication appearing in the window, the search will be terminated before the neighbor (left/right) medication mention.
3. Search will be terminated by some punctuation marks (eg, period '.', semicolon ';', or comma ',') if it is necessary.
4. If the constituents within the search text imply multiple dosages, modes, frequencies, durations, or reasons, the closer the constituent is to the medication mention, the higher the confidence score is assigned to it.

Finally, given a discharge summary, the system would output a list of medication entries. The format for each medication entry extracted from the text is described as follows (a medication entry sample is available in a *JAMIA* online data supplement available at http://jamia.bmj.com):

medication name & offset || dosage & offset || mode & offset || frequency & offset || duration & offset || reason & offset || list/narrative.

RESULTS AND DISCUSSION

Evaluation metrics used

For the I2B2 medication extraction task, the organizers used two kinds of evaluation metrics: exact matching at the term level and inexact matching at the token level. Given aligned system output, two kinds of evaluation measures were performed: horizontal (medication entry layer) and vertical (medication constituent category layer). The organizers performed evaluation at two different levels of granularity: (1) Patient record level: first micro-average over all entries in a single record, and then macro-average over all the records in the system output. (2) System level: micro-average over all entries in the system output. More detailed information about system evaluation metrics can be found in the overview paper of the i2b2 NLP challenge.[1]

Results

The competition results were graded versus the gold standard using an F-measure micro-average score. We submitted two sets of results. Table 4 summarizes our system performance for medication extraction against the manually annotated gold standard. Our system performed well and the best results achieved an F-micro score of 80% on exact matching (run 2) and 81% on inexact matching (run 2) for the system-level horizontal evaluation. The slight difference between exact matching and inexact matching results suggests that most of the false-positive instances in exact matching do result from mismatching rather than from partial matching.

For the system-level vertical evaluation, the best categories were medication names and mode/route with the F-micro of over 0.848. It suggests that the term lists collected for these two categories were relatively complete and included most of the mention occurrence cases in the test data. Moreover, our token-based

Table 4 The performance for medication extraction on the annotated test data

			Run 1 (F-measure)		Run 2 (F-measure)	
			Exact	Inexact	Exact	Inexact
Horizontal	med-entry	System-level	0.7955	0.8111	0.7955	0.8116
		Patient-level	0.7845	0.7973	0.7846	0.7977
Vertical	med name	System-level	0.8589	0.8904	0.8589	0.8904
		Patient-level	0.8593	0.8876	0.8593	0.8876
	dosage	System-level	0.8037	0.8596	0.8037	0.8596
		Patient-level	0.7942	0.8496	0.7942	0.8496
	mode	System-level	0.8488	0.8543	0.8488	0.8543
		Patient-level	0.8409	0.8489	0.8409	0.8489
	frequency	System-level	0.8170	0.8299	0.8165	0.8322
		Patient-level	0.8063	0.8300	0.8059	0.8329
	duration	System-level	0.400	0.4494	0.3987	0.4489
		Patient-level	0.3451	0.3934	0.3387	0.3866
	reason	System-level	0.2294	0.2439	0.2294	0.2439
		Patient-level	0.2202	0.2520	0.2202	0.2520

morphological rules matching method had been proved relatively successful, which also achieved an F-micro of over 0.80 in both dosage and frequency extraction. It indicates that the token-based rule method was an effective approach to resolving the morphological variation problem. However, duration and reason extraction performed relatively poorly with F-micro scores of only 0.4 and 0.23, respectively. The main reason for that was that, for the challenge, we went with short duration and reason term lists that were unable to have a good handle on the widely varying number of duration and reason terms involved for this challenge.

Discussion

There are several possibilities for further improvement of our rule-based approach. Based on the error analysis of false negative results, the medication information integration component would need to be enhanced by developing a new integration strategy to detect multiple entries from a simple medication mention (eg, 'OTC OMEPRAZOLE 20 MG PO ONE TAB QD' → 'otc omeprazole & 20 mg & po', 'otc omeprazole & one tab & qd') or by creating text patterns to identify the cases in which multiple events take place (eg, 'We increase his vitamin C to 20 mg' → 'vitamin C' (stop event), 'vitamin C & 20 mg' (start event)).

The error analysis of false-positive results revealed that our string-based lookup approach had difficulties in several complicated term-matching situations: (1) the lack of standardization of medication names; (2) long and descriptive naming convention; (3) conjunctive and disconjunctive structure; (4) causal naming; (5) misspellings. More regular expression rules are thus required to cope with the above complex situations with the addition of a spell-check function. More error analysis details can be found in the *JAMIA* online data supplement available at http://jamia.bmj.com.

CONCLUSIONS

In this paper, we reported on our approach for the i2b2 medication extraction challenge. We developed a relatively simple rule-based approach with manually curated feature term lists and token-based regular expression rules. As indicated by its sixth place ranking in exact matches and fourth in inexact matches, this approach performed reasonably well for this multiple entities identification task with few annotated data available, and was competitive with the approaches by other participant teams using pre-existing domain-specific tools and resources. However, poor performance in duration and reason extraction suggests the addition of machine learning algorithms trained on an appropriate size of the annotated data has the potential for enhancing the performance of the system, particularly for situations involving the construction of large term vocabulary resources with wide variants of the terms.

Acknowledgments The author would like to thank i2b2 for organizing the 2009 i2b2 medication extraction challenge and providing this research opportunity.

Funding 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, Cadga E. Extracting medication information from clinical text. *J Am Med Inform Assoc* 2010;**17**:514—8.
2. GENIA Tagger. http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/ (accessed 10 Dec 2009).
3. **UMLS Knowledge Base.** http://www.nlm.nih.gov/research/umls (accessed 10 Dec 2009).
4. **SNOMED-CT.** http://www.ihtsdo.org/snomed-ct/ (accessed 10 Dec 2009).
5. **Chapman W,** Bridewell W, Hanbury P, *et al*. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 2001;**34**:301—10.