High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge

Jon Patrick, Min Li

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

Health Information Technology Research Laboratory, School of IT, Faculty of Engineering and IT, the University of Sydney, Sydney, Australia

Correspondence to
Dr Jon Patrick, School of IT, Faculty of Engineering and IT, the University of Sydney, Sydney, NSW 2006, Australia; jonpat@it.usyd.edu.au

Received 20 February 2010
Accepted 25 June 2010

ABSTRACT

Objective Medication information comprises a most valuable source of data in clinical records. This paper describes use of a cascade of machine learners that automatically extract medication information from clinical records.

Design Authors developed a novel supervised learning model that incorporates two machine learning algorithms and several rule-based engines.

Measurements Evaluation of each step included precision, recall and F-measure metrics. The final outputs of the system were scored using the i2b2 workshop evaluation metrics, including strict and relaxed matching with a gold standard.

Results Evaluation results showed greater than 90% accuracy on five out of seven entities in the name entity recognition task, and an F-measure greater than 95% on the relationship classification task. The strict micro averaged F-measure for the system output achieved best submitted performance of the competition, at 85.65%.

Limitations Clinical staff will only use practical processing systems if they have confidence in their reliability. Authors estimate that an acceptable accuracy for a such a working system should be approximately 95%. This leaves a significant performance gap of 5 to 10% from the current processing capabilities.

Conclusion A multistage method with mixed computational strategies using a combination of rule-based classifiers and statistical classifiers seems to provide a near-optimal strategy for automated extraction of medication information from clinical records.

Many of the potential benefits of the electronic medical record (EMR) rely significantly on our ability to automatically process the free-text content in the EMR. To understand the limitations and difficulties of exploiting the EMR we have designed an information extraction engine to identify medication events within patient discharge summaries, as specified by the i2b2 medication extraction shared task.

BACKGROUND

In the literature, a number of studies have focused on the drug recognition task, and most of them have used a rule-based approach. For example, the earliest work we have found is the CLAPIT natural language processing system.[1] In this work, the extraction targets were drug and dosage phrases and the researchers achieved an F-score of approximately 86.7% in exact matches. In the work of Sirohi and Peissig,[2] the drug extraction relied on a carefully selected drug lexicon. They reported a high recall of 95.8%, but lower precision of 54.6%. A recent study by Gold *et al*[3] documented a rule and lexicon-based approach for extracting medication information, which achieved an F-score of nearly 87.9% for the drug extraction.

Compared with the previous studies, the i2b2 workshop definition for the medication information is much broader, and includes the medication name, dosage, mode of administration, frequency, duration and its offset, reason, and their offsets in each case, and found in list/narrative of the text. Unlike the systems described in the literature, we used a supervised machine learning-based approach which integrated rule-based methods. The final results show this approach is better than the rule-based approach which was adopted by other teams in the challenge.

METHODS

Challenge requirements

For the 2009 i2b2 challenge, the main objective is to provide accurate and comprehensive information about the medications a patient has been administered based on the evidence appearing in the textual records. For each medication entry, the information that needed to be extracted included medication name, dosage, mode of administration, frequency, duration, reason, and context.[4] Multiple medication entries should be generated if the medication is a change of dosage or multiple dosages, modes, frequencies, durations and reasons (see figure 1 in online appendix at http://jamia.bmj.com).

Corpus description

The challenge corpus is composed of 1249 patient discharge summaries provided by Partners Healthcare. It was split into training (696 records) and test (533 records) sets. From the training set, we selected 145 of the longest records, because the longer records always contain more medication information than the shorter records. Two team members, one researcher and one physician, manually annotated these records. A sequential annotation strategy was adopted, in which the physician annotated the records first and the researcher revised them.

The classification strategy

Our approach is a cascade of machine learning classifiers, which integrated some rule-based methods. The two machine learners adopted were conditional random fields (CRF)[5] and support vector machines (SVM).[6] The first machine learner was used to identify named entities, while the

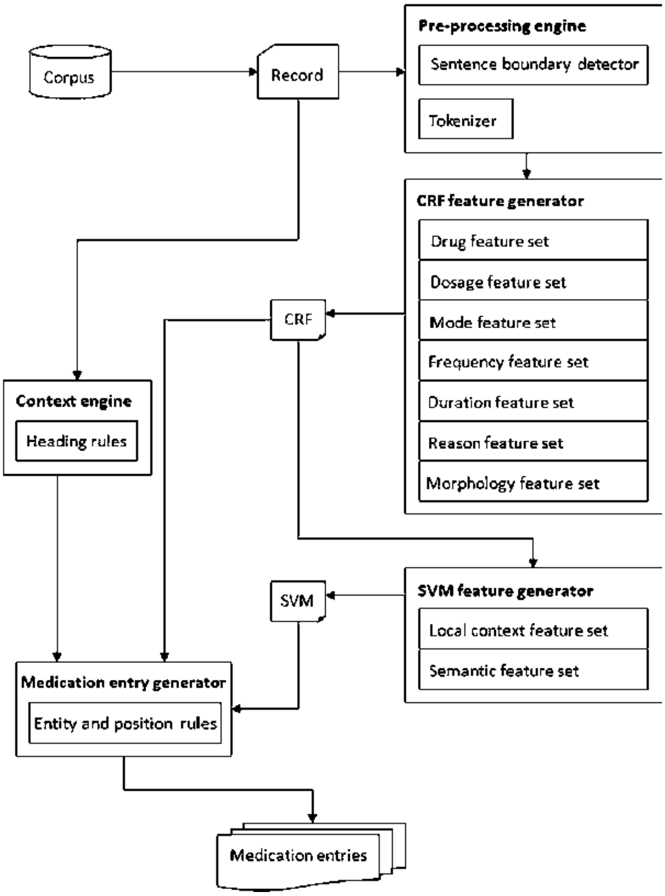

Figure 2 Medication extraction system architecture diagram. CRF, conditional random fields; SVM, support vector machines.

second was utilized to classify the relationships between two entities.[7]

Figure 2 shows the system architecture. From this diagram, every record in the corpus moves through the pre-processing engine, which includes a sentence boundary detector and a tokenizer. The sentence boundary detector relies heavily on a precompiled lexicon. This lexicon was generated from the training set which stored the abbreviations embedded with a period, such as 'q.i.d.', 'Dr.', 'h.s.', 'iv.' etc. If there was a period/ exclamation/ question mark at the end of the current token, and this token was not included in the lexicon, and the next token was capitalized, then it ended a sentence. In this way, the records were split into sentences. The tokenizer was a white space tokenizer specified by the challenge.

After this preprocessing, seven CRF feature sets were prepared to train the CRF to identify the named entities, which were sent to the SVM feature generator. The feature sets were composed of a local context feature set and a semantic feature set. These feature sets were used to train the SVM to classify the relationship between two entities.

The context engine was a rule-based engine, which identified whether or not the medication was under a medication heading in the report. Finally, when the results from the CRF, SVM and context engine were ready, the medication entry generator generated the medication entities according to the format defined for the challenge.

The CRF experiment for named entity recognition

The purpose of this experiment is to extract the medication, dosage, mode, frequency, duration and reason entities. Six feature sets and a five-word window around each word constituted the feature model. The feature sets were:

1. Drug feature set: a combination of the results from the drug lexicon verification, drug gazetteer lookup, drug pattern mapping engine, and negation engine.
2. Dosage feature set: derived from the dosage gazetteer lookup and the dosage mapping engine.
3. Mode feature set: derived from a mode gazetteer lookup. The mode gazetteer was generated from the training set, with 60 names.
4. Frequency/duration feature set: the frequency/duration was generated with the same approach as the dosage feature set.
5. Reason feature set: a combination of the results from 'text to SNOMED CT' conversion process,[8] with compilation of a reason gazetteer and reason prediction engine, but using only the 'clinical finding' category in SNOMED CT.
6. Morphology feature set: the prefixes and suffixes of words.
7. Five-word context window: the selection of window size was a separate process. Nine different context word window sizes (2—10) were experimented with, the results showing that the five-word window size was optimal.

Many other features were experimented with, such as the part of speech features, orthographic features, and medical category for each word. However, the best performance was obtained from the seven feature sets (see appendix table 1, available online only at http://jamia.bmj.com).

The SVM experiment for relationship identification

In order to connect the six entities to generate full medication entries, an SVM was used to classify the relationships between entity pairs of a medication and its related five entities once the name entity recognition (NER) task was completed (see figure 3 and table 2, both available in online only appendix at http://jamia.bmj.com). Six feature sets for computing valid relationships between entity pairs were used to train the SVM, consisting of local context features and semantic features:

1. Local context features:
   a. three words before and after the first entity
   b. three words before and after the second entity
   c. words between the two entities
   d. words inside of each entity.
2. Semantic features:
   a. the types of the two entities determined by the CRF classifier
   b. the entity types between the two entities.

The feature selection mechanism is the same as for NER feature selection.

Context engine

The context engine identifies the medication entry under the special section headings, such as 'medications on admission:', 'discharge medications:', etc, or in the narrative part of the clinical record. There are four stages to identifying whether the medication entry is in these sections or not, which result in finding the span between the medication heading and the next heading, and if the medication event appears in this span, it should be extracted as valid.

Medication entry generator

The medication entry generator is the final step in this system and is responsible for assembling all the components into the final medication event entries based on having established their relationships. The results from the previous steps are used here, namely CRF, SVM and context engine. Two different strategies are required in this step (see figure 4, available in online only appendix at http://jamia.bmj.com):

1. Using the SVM results to identify the medication entries:
   The context value (list/narrative) is determined by the context engine. The algorithm which is used to build medication entries is based on the position rule of each entity and the total number of elements in each entity type. Seven different cases can be defined.
   a. In the first case the total number of elements in each entity is less than two so two medication entries should be generated, since the keywords like 'increase', 'decrease', always means a change of dosage.
   b. Otherwise, only one medication entry should be generated.
   c. Two dependent frequencies should be connected together, thus, only one medication entry should be generated (see figure 5, available in online only appendix at http://jamia.bmj.com).
   d. Otherwise two separate medication entries are required.
2. The position rule is adopted in the three cases where (see figure 6, available in online only appendix at http://jamia.bmj.com):
   a. The total number of elements for any two entity types is greater than 1 and the remaining elements less than 2.
   b. The total number of elements for any three entity types is greater than 1 and the remaining elements less than 2.
   c. The total number of elements for each entity type is greater than 1.
3. Individual medication generation:
   If the medication in the clinical notes does not have any relationships with other entity types, it will be missing from the SVM result. Consequently, this medication should be withdrawn from the CRF results and an individual medication entry generated for it. The value for the context (list/narrative) also comes from the context engine, as in the previous step.

RESULTS

Table 3 summarizes our system performance for medication extraction on the 251 test records. 'Patient level' in column 3 means every token in the entity is evaluated individually while the whole entity will be evaluated at the 'system level'. In the patient-level case, all notes, regardless of length will have equal weighting for the final F-score (macro-averaging over the notes). On the other hand, with system level performance, a note with more medication entries would have more effect on the final score (micro-averaging over all entries).

DISCUSSION

Conditional random fields experiment results analysis

Reason and duration are the most difficult entities to recognize (their average F-score is approximately 50%, while the mode, dosage and frequency perform best with an average F-score greater than 90%). This occurs because of the much smaller frequencies for the reason and duration than the other four entity types.

Reason extraction, in our system, depends heavily on the finding category in SNOMED CT and the performance of the text to SNOMED CT process. However, the finding category is not a good match to the reason entities in the clinical notes, due to the many varied ways reason can be expressed that will not exist in SNOMED CT, as well as the manner in which reasons can be ambiguously expressed.

Support vector machine experiment results analysis

The SVM experiments determined relationships between entities. If medication and its reason is in a single sentence, focus is on the pairs found in one sentence. However, sometimes medication and its reason could be distributed across two contiguous sentences. A high performance is achieved in which the F-score for the 'has relation' set of the single sentence level is 97.53%, while 95.91% is achieved in the paired sentence level, indicating little if any systematic errors. The performance for the single sentence level is higher than the two sentence level due to the greater number of combinations of entity pairs contained in paired sentences. This is the reason for only considering the relationship of the 'medication—reason' pair in the paired sentence model as a supplement to the related pairs gathered in the single sentence level. The remainder of the relationships identified in the paired sentences are not used in the system.

Final results analysis

Due to the errors in the NER, relationship classification and medication entry generator, the final F-scores for each entity type are lower than in the NER processing.

The final score for the exact match medication entry was 85.65%. The main reason for the performance decrease in dosage, mode, frequency, duration and reason is because of the

Table 3 Final evaluation scores for NER and relationship classification (reference to appendix table 3, available online only)

			Exact			Inexact		
			Precision	Recall	F-score	Precision	Recall	F-score
Horizontal	Med-entry	System-level	0.8962	0.8202	0.8565	0.9033	0.8010	0.8490
		Patient-level	0.8962	0.8138	0.8488	0.9028	0.7938	0.8390
Vertical	Drug	System-level	0.9085	0.8598	0.8835	0.9289	0.8777	0.9026
		Patient-level	0.9087	0.8588	0.8798	0.9322	0.8765	0.8996
	Dosage	System-level	0.9189	0.8678	0.8926	0.9317	0.8860	0.9083
		Patient-level	0.9125	0.8646	0.8815	0.9254	0.8803	0.8959
	Mode	System-level	0.9073	0.8915	0.8994	0.9090	0.8780	0.8932
		Patient-level	0.9038	0.8894	0.8906	0.9054	0.8836	0.8868
	Frequency	System-level	0.9142	0.8795	0.8965	0.9236	0.8346	0.8768
		Patient-level	0.9144	0.8774	0.8912	0.9281	0.8509	0.8799
	Duration	System-level	0.5604	0.3709	0.4464	0.6380	0.4092	0.4986
		Patient-level	0.4962	0.3805	0.4104	0.5634	0.4489	0.4700
	Reason	System-level	0.6687	0.3319	0.4436	0.6654	0.3231	0.4350
		Patient-level	0.5873	0.3571	0.4149	0.6106	0.3695	0.4234

NER, name entity recognition.

low recall for medication in the NER (computed using CRF). Another factor is the low performance of reason extraction by the NER. The frequency of appearance of multiple reasons is relatively high, and the multiple reasons should be used to construct multiple medication entries (see example 2 in appendix figure 1, available online only). In this way, the loss in reason recognition has led to the decrease in recall of all other entity types and the medication event.

CONCLUSIONS

We have introduced a complex machine learning model that was designed to participate in the 2009 i2b2 medication extraction challenge. This model was based on a cascaded approach, which integrated CRF, SVM and several rule-based engines. The final results demonstrated this system performance was optimal relative to other participants (see table 4, available in online only appendix at http://jamia.bmj.com).

From the NER experiment, it is obvious that duration and reason were the two weakest parts and need to be improved.

Precision and recall are both important for clinical staff who will only use practical processing systems if they have confidence they will return all and only the wanted results. The experience we have had in a variety of hospital settings leads us to believe that an acceptable level of performance would have to be minimally F>90.0 and perhaps F>95.0. As our current processing capabilities are just below 86% this leaves a significant gap of 5–10% to reach an acceptable level of accuracy.

Acknowledgments The authors would like to acknowledge the contribution of Dr Stephen Crawshaw, Dr Yefeng Wang and other members in the Health Information Technologies Research Laboratory. De-identified clinical records used in this research were provided by the i2b2 National Center for Biomedical Computing funded by U54LM008748 and were originally prepared for the Shared Tasks for Challenges in NLP for Clinical Data organized by Dr Ozlem Uzuner, i2b2 and SUNY.

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. **Evans DA,** Brownlow ND, Hersh WR, *et al*. Automating concept identification in the electonic meidcal record: an experiment in extracting doseage information. *American Medical Informatics Association.* Washington, DC, October 26–30, 1996. Symposium Proceedings: 388–92.
2. **Sirohi E,** Peissig P. Study of effect of drug lexicons on medication extraction from electronic medical records. *Pacific Symposium on Biocomputing*, Hawaii, January 4–8, 2005. 2005;**10**:308–18.
3. **Gold S,** Elhadad N, Zhu X, *et al*. Extracting structured medication event information from discharge summaries. *American Medical Informatics Association*. Washington, DC, November 8–12, 2008. Symposium Proceedings 2008;237–41.
4. **Uzuner Ö,** Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc* 2010;**17**:514–8.
5. **CRF++.** Yet Another CRF toolkit. Software available at: http://crfpp.sourceforge.net/ (accessed 8 Aug 2010).
6. **Cristianini N,** Shawe-Taylor J. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge, UK: Cambridge University Press, 2000.
7. **Wang Y,** Patrick J. Evaluating linguistic features for clinical relation extraction. *Proceedings of the Conference of the Pacific Association for Computational Linguistics (PACLING)*, Sapporo, Japan, September 1–4, 2009:349–55.
8. **Patrick J,** Wang Y, Budd P. An automated system for conversion of clinical notes into SNOMED clinical terminology. *Proceedings of the 5th Australasian Symposium on ACSW Frontiers*, Ballarat, Australia, 2007;**68**:219–26.