A Cascade Approach to Extracting Medication Events

Jon Patrick
School of IT
The University of Sydney
Sydney, NSW 2006, Australia
jonpat@it.usyd.edu.au

Min Li
School of IT
The University of Sydney
Sydney, NSW 2006, Australia
mili9528@uni.sydney.edu.au

Abstract

Information Extraction, from the electronic clinical record is a comparatively new topic for computational linguists. In order to utilize the records to improve the efficiency and quality of health care, the knowledge content should be automatically encoded; however this poses a number of challenges for Natural Language Processing (NLP). In this paper, we present a cascade approach to discover the medication-related information (MEDICATION, DOSAGE, MODE, FREQUENCY, DURATION, REASON, and CONTEXT) from narrative patient records. The prototype of this system was used to participate the i2b2 2009 medication extraction challenge. The results show better than 90% accuracy on 5 out of 7 entities used in the study.

1 Introduction

Electronic records are widely used in the health care domain since we believe they can provide more advantages than the traditional paper record (Sujansky, 1998). However, the value of electronic clinical records depends significantly on our ability to discover and utilize the specific content found in them. Once this content can be detected, the potential benefits for individual clinicians and healthcare organizations are considerable.

In this study we focus on discharge summaries, which have their own challenges. This kind of clinical record includes several sections. The average word count of these reports is around 1500 words per record. This paper presents a method to extract all the medication related information, and connect the relative entities together to build medication entries using a cascaded approach based on two machine learners.

2 Related Work

In this paper, we focus on both NER and RC tasks to extract the medications and their related information (DOSAGE, MODE, FREQUENCY, DURATION, REASON, and CONTEXT) from free-text clinical records. At this time, it's difficult to compare our system with other systems which participated the i2b2 2009 medication extraction challenge, since these publications are unavailable now. Consequently, we can only compare our system with some similar studies in the literature. In the previous work, only three published studies address this issue (see the performance comparison in the final section) and these studies do not have a comprehensive and precise definition of medication information. The closest research for medication event extraction relies on parsing rules written as a set of regular expressions and a user-configurable drug lexicon. It includes the event for DRUG, DOSAGE, ROUTE, FREQUENCY, CONTEXT and NECESSITY (Gold et al. 2008). The basic work flow for their system starts by discovering drug names based on a drug dictionary, and the rest of the process uses the MERKI parser.

The CLARIT NLP system (Evans et al. 1996) can extract DRUG-DOSAGE information from clinical narratives. This system is based on the rule-based method and five main steps are included, such as tokenization, stemming, syntactic category assignment, semantic category assignment and pattern matching.

Another system focuses on the drug extraction only and is based on a drug lexicon (Sirohi and Peissig 2005). This study demonstrates that high precision and recall for medication extraction from clinical records can be obtained by using a carefully selected drug lexicon.

Comparing these three medication extraction systems, a different approach is adopted in our work. Our medication event system is based on the combination of a machine learner approach and rule based approach. Two machine learners were used, namely the conditional random field (CRF) and support vector machine (SVM). Moreover, a broader definition for a medication event is considered, especially the REASON for the medication which hasn't been studied in previous research. Furthermore, the medication information in our training and test set is much larger than prior studies.

3 Methodology

There are four main steps in our methodology:

1. Definition of the information to be extracted.
2. Preparing data for training and testing.
3. Using natural language processing technologies to build a medication event extraction system.
4. Passing the test data to the system and evaluation of the final result.

3.1 Extraction Definition

Our goal is to provide accurate, comprehensive information about the medications a patient has been administered based on the evidence appearing in the textual records. For each medication entry, the following information needs to be extracted: Medication, dosage, mode, frequency, duration, reason, and context.

Multiple medication entries should be generated if the MEDICATION has the changes for DOSAGE or multiple DOSAGEs, MODEs, FREQUENCYs, DURATIONs and REASONs.

3.2 Data Preparation

One hundred and sixty clinical records were prepared for training (130 records) and testing (30 records). One physician and one researcher created the gold standard annotations by sequential annotation: the physician annotated the records first and his results were given to the researcher to revise. The annotation process took approximately 1.5 hours per record due to the length of clinical records.

4 Medication Event Extraction System Architecture

The basic strategy for the medication event extraction system is to: ① use CRF to identify the entities, ② build pairs for each medication relationship (only consider DRUG and its related entity, since the whole related entities, such as DOSAGE, FREQUENTY, etc., could be further connected based on the DRUG), ③ classify the binary relationships by SVM, ④ generate medication entries based on the results from the CRF and SVM. Figure 1 demonstrates the detailed system architecture, which includes the following processing stages:

I. Sentence Spitting

Split the clinical records into individual sentences.

II. Tokenization

Each sentence is split into tokens their position and extent in the text.

III. CRF Feature Builder

Seven feature sets were prepared in this stage, to be used in the CRF training. They are DRUG, DOSAGE, MODE, FREQUENCY, DURATION, REASON, and morphology feature sets.

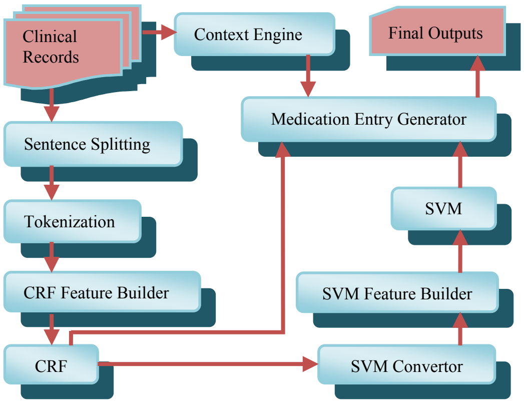

Figure 1. NER and RC System Architecture.

IV. CRF Model Building and Classification

The CRF feature builder generated the features for the CRF machine learner. The context window for the CRF was set to be five words.

V. CRF Model Building and Classification

The CRF results were converted into SVM input features by the SVM Convertor. There are two kinds of SVM input generated here:

1. Unigram Sentences

Each pair of medication elements at the unigram sentence level is used to build an SVM training record.

2. Sentence Pairs

Sometimes MEDICATION and its REASON could be across two sentences. Like the mechanism to generate the unigram sentence input, medication pairs are also built at the sentence pair level.

VI. SVM Feature Builder

Six Features are generated based on the output from the SVM Convertor to classify the relationships:

1. Three words before and after the first entity.
2. Three words before and after the second entity.
3. Words between the two entities.
4. Words inside of each entity.
5. The types of the two entities determined by the CRF classifier.
6. The entities types between the two entities.

VII. SVM Model Building and Classification

The features which were generated in the previous step were passed to the SVM to build the model and classify the relationships between medication pairs for the test set.

VIII. CONTEXT Identification

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. The performance is discussed in the next section.

IX. Medication Entry Generation

The medication entry generator is the final step in this system which 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 stages are involved in this step:

(a) Using the SVM results to identify the medication entries. The CONTEXT value (list/narrative) comes from 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 each entity type. It can be divided into several cases.

(b) If the medication in the clinical notes doesn't 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.

5 Results and Discussion

In this section, the experiment results for NER, RC, CONTEXT engine and the final output for the test set is presented and discussed.

5.1 NER(CRF) Experiment

The main purpose of this experiment is to extract the MEDICATION, DOSAGE, MODE, FREQUENCY, DURATION and REASON from the clinical records. Table 1 demonstrates the performances for exact match by using the 7 feature sets. The number in the bracket is the baseline, which use the bag of words as the only feature set. The baseline shows the extraction for REASON and DURATION are the most difficult entities to recognise (their average F-score is about 50%, while the MODE, DOSAGE and FREQUENCY perform best with an average F-score greater than 92%).

Entity Type	Training	Test	Recall (Baseline)	Precision (Baseline)	F-Score (Baseline)
Overall	**17337**	**5296**	**88.82% (80.25%)**	**92.89% (93.49%)**	**90.81% (86.36%)**
MEDICATION	6576	1940	91.44% (76.34%)	91.35% (91.87%)	91.40% (83.39%)
DOSAGE	3352	1076	93.49% (88.66%)	96.36% (95.69%)	94.91% (92.04%)
MODE	2537	796	94.60% (91.21%)	95.92% (96.93%)	95.26% (93.98%)
FREQUENCY	3180	1020	93.24% (90.26%)	96.26% (95.74%)	94.72% (92.94%)
DURATION	366	104	51.92% (41.35%)	80.60% (79.63%)	63.16% (54.43%)
REASON	1326	360	46.11% (34.72%)	69.75% (72.67%)	55.52% (46.99%)

Table 1. Best scores and baseline scores from CRF of NER

It is worth pointing out many other features were experimented with during the system implementation, such as the medical category for each word, whether the word is capitalized, in lower case or upper case, etc. However, the best performance is obtained from the 7 feature set. The feature selection process is that:

In the first place, all features were gathered together to train the model and predict the results. Sequentially, the performance of this experiment was recorded. Next, we did a set of experiments to remove every feature from the whole features one by one, and then train the related model. After that, predict the results and record the performance. Finally, these performances were compared with the performance in the first step to see whether the removed feature decreased in the F-score. If it did, this feature would be useful. Else, it was useless.

The performances for the REASON and DURATION are still the lowest, but the F-scores are approximately 10% higher than the baseline. This is because:

1. The frequencies for the REASON and DURATION are much smaller than the other four entity types.
2. For the DURATION entities, the rule based regular expression can match other non-medication terms. Also, there are some DURATION terms that can't be discovered by our rules.
3. REASON extraction depends highly on the Finding category in SNOMED CT and the performance of TTSCT (Patrick et al. 2007). However, the Finding category cannot be well-matched to the REASON entities in the clinical notes, due to the many varied ways REASON can be represented which may not exist in the SNOMED CT, and as well the REASONs that are ambiguously expressed. Another limitation is the performance of TTSCT. Consequently, these issues lead to low performance on REASON, and the F-score of DURATION (63.16%) is higher than the REASON (55.52%) even though the frequency of DURATION is smaller than REASON (104 and 360 respectively).

Compared to the baseline, the F-scores for the MODE, DOSAGE and FREQUENCY were only improved by about 2%. The first reason is that the performance of the baseline is already very high (around 90%). Secondly, the regular expressions and gazetteers cannot capture all the different ways to present these three entity types. Approximately 8% improvement in the MEDICATION extraction is obtained in the system, since the medication lexica were used in the system. The errors come from:

1. Misspelling of drug names, such as "nitroglycerin"
2. Drug names used in other contexts, such as the "coumadin" in the "Coumadin Clinic" phrase.
3. The drug allergies detector cannot cover all situations.

Overall, the system scored of 90.81% on the NER task.

5.2 Relationship Classification Experiment

The support vector machine is used to classify the relationships between the medication pairs (see section 2). The feature sets used are discussed in the previous section. Meanwhile, the feature selection mechanism is same as the NER feature selection, which was introduced in the previous sub-section. For comparison, the baseline only uses three of the whole feature sets, namely, No.1, 2 and 4 in the SVM feature sets. Two experiments were conducted (the unigram sentence level and sentence pair level) for the baseline and subsequent solutions.

Relation Type	Total Number	Recall (Baseline)	Precision (Baseline)	F-Score (Baseline)
HAS RELATIONSHIP (unigram)	**3373**	**98.89% (82.69%)**	**97.90% (61.26%)**	**98.39% (70.38%)**
NO RELATIONSHIP (unigram)	24765	99.71% (92.96%)	99.85% (97.55%)	99.78% (95.20%)
HAS RELATIONSHIP (sentence pair)	**7030**	**97.06% (82.47%)**	**95.89% (63.53%)**	**96.47% (71.77%)**
NO RELATIONSHIP (sentence pair)	48162	99.40% (93.19%)	99.58% (97.36%)	99.49% (95.23%)

Table 2. Best scores and baseline scores from SVM of RC

The baseline F-score for the HAS RELATIONSHIP set of the unigram sentence level is 70.38% and 95.20% in the NO RELATIONSHIP set. The difference can be attributed to the fact that the total number of the NO RELATIONSHIP set is 7 times larger than the HAS RELATIONSHIP set. However, the performance in “has relation” is more important, since the generation of medication entries is based on the pairs which have the relationship correctly identified.

A high performance is achieved in which the F-score for the “has relation” set of the unigram sentence level is 98.39%, while 96.47% is achieved in the bigram sentence level indicating little if any systematic errors.

5.3 CONTEXT Engine Evaluation

The CONTEXT engine was adopted to discover the span of the medication list (the span between the medication heading and the next following heading). The rules which are used in the engine are based on the medication headings in the training set. Table 3 shows the performance of the test set for the CONTEXT engine.

Entity Type	Training	Testing	Recall	Precision	F-Score
Heading pairs	166	51	94.44%	100.00%	97.14%

Table 3. System scores from SVM for determining Context.

An F-score of 97.14% was achieved with the CONTEXT engine.

5.4 Final Output Evaluation

The final evaluation tool used here is released from i2b2 National Center. 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 scores for the medication event are between 86.23% and 88.16% (see table 4). The main reason for performance decrease in DOSAGE, MODE, FREQUENCY, DURATION and REASON is because the low recall for the MEDICATION in the NER (computed using CRF). If these medications related entities were extracted without the MEDICATION, these entities could not be connected into medication entries, which make them meaningless in the final output. 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. In this way, the loss in REASON recognition would lead to the decrease in recall of all other entity types and the medication event.

Type	Token Level F-Sore	Entity F-Score
Medication Entry	87.33%	88.16%

Table 4. Final evaluation scores for Medication Entry.

6 Conclusion

In this paper, a high accuracy and comprehensive medication event extraction system is presented. Compared to the three similar systems (see section 2), a better performance is achieved here, even through these systems have a narrower definition for medication event and a different evaluation metric. For example, the F-score of MEDICATION in Sirohi’s system is 69.55%, whereas our system achieves 91.40%. As well, the F-score of the exact match for DRUG-DOSAGE event in the Evans’s system is 86.76% and 87.92% is obtained in Gold’s system for the MEDICATION in their medication event. In contrast, the MEDICATION in the medication event of our system achieves an F-score of 89.16%~90.93%.

In future work, DURATION and REASON are the two main entities that need to be improved. One possible solution is to use the relationship between the medication and its corresponding diseases or symptoms to improve the REASON extraction. As to DURATION, increasing the training set to obtain more examples is probably the best strategy.

Acknowledgments

We would like to acknowledge the contribution of Stephen Crawshaw, Yefeng Wang and other members in the Health Information Technologies Research Laboratory.

Deidentified 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.

References

David A. Evans, Nicholas D. Brownlowt, William R. Hersh, and Emily M. Campbell. 1996. Automating Concept Identification in the Electionic Meidcal Record: An Experiment in Extracting Doseage Information. *AMIA 1996 Symposium Proceedings*, 388-392

Sigfried Gold, Noémie Elhadad. and Xinxin Zhu. 2008. Extracting Structured Medication Event Inforamtion from Dicharge Summaries, *AMIA 2008 Symposium Proceedings*, 237

Jon Patrick, Yefeng Wang and Peter Budd. 2007. An automated system for conversion of clinical notes into SNOMED clinical terminology, *in Proc. 5rd Australasian symposium on ACSW frontiers*, 68: 219-226.

E. Sirohi, and P. Peissig. 2005. Study of Effect of Drug Lexicons on Medication Extraction From Electronic Medical Records. *Pacifi Symposium on Biocomputing.* 10: 308-318

Walter V. Sujansky. 1998. The benefits and challenges of an electronic medical record: much more than a "word-processed" patient chart. *West J Med*, 169(3):176-83.