Journal of Biomedical Informatics 72 (2017) 60–66

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Prescription extraction using CRFs and word embeddings

Carson Tao [a,*], Michele Filannino [b], Özlem Uzuner [b]

[a] Department of Information Science, State University of New York at Albany, NY, USA
[b] Department of Computer Science, State University of New York at Albany, NY, USA

ARTICLE INFO

Article history:
Received 11 January 2017
Revised 23 June 2017
Accepted 3 July 2017
Available online 4 July 2017

Keywords:
NLP
Machine learning
Word embeddings
CRFs
Prescription extraction

ABSTRACT

In medical practices, doctors detail patients' care plan via discharge summaries written in the form of unstructured free texts, which among the others contain medication names and prescription information. Extracting prescriptions from discharge summaries is challenging due to the way these documents are written. Handwritten rules and medical gazetteers have proven to be useful for this purpose but come with limitations on performance, scalability, and generalizability. We instead present a machine learning approach to extract and organize medication names and prescription information into individual entries. Our approach utilizes word embeddings and tackles the task in two extraction steps, both of which are treated as sequence labeling problems. When evaluated on the 2009 i2b2 Challenge official benchmark set, the proposed approach achieves a horizontal phrase-level F1-measure of 0.864, which to the best of our knowledge represents an improvement over the current state-of-the-art.

1. Introduction

In medical practices, doctors detail patients' care plan in unstructured free texts. These documents contain medication names and prescription information, which are important components of patients' overall care.

Extracting medication names and other prescription information from discharge summaries is challenging due to the way these documents are written. While highly readable to those with medical background, such documents are not intended to be digested by computers. The presence of different expressions conveying the same information, medical acronyms, misspellings, and ambiguous terminologies makes the automatic analysis of these documents difficult.

For example, consider the following excerpts: (1) The doctor prescribed him 325 mg Aspirin p.o. 4x/day for 2 weeks as needed for inflammation. (2) We gave her a seven-day course of 200 mg Cefpodoxime q.h.s. for bronchitis, which was taken through mouth. From both excerpts, we want to extract mentioned medication names along with information related to their dosage, mode of administration, frequency, duration, and medical reason for prescription. We refer to this task as medication information extraction, where medication names, dosages, modes, frequencies, durations, and reasons are medication entities. Medication entities corresponding to the above examples are demonstrated in Table 1 below.

We then group medication entities together through *relation extraction* to create *medication entries*, which link medications to their signature information and constitute the final output. Medication information extraction and relation extraction collectively make up what we refer to as *prescription extraction.*

In this paper, we present a system for automatic prescription extraction from unstructured discharge summaries. We treat prescription extraction as a two-step sequence labeling task: we first apply Conditional Random Fields (CRFs) with word embeddings to extract medication information; we then tackle relation extraction as a second sequence labeling task. We evaluated our system against the i2b2 2009 official benchmark set on medication entries, achieving a horizontal phrase-level F1-measure of 0.864 (see section 4.4 for a description of the evaluation metrics). The proposed system achieves a significantly higher overall performance than the current state-of-the-art system.

2. Related work

2.1. Medication extraction systems

MedLEE [1] is one of the earliest medication extraction systems, built with the purpose of extracting, structuring, and encoding clinical information within free text patient reports. MedLEE extracts medication information using hand-written rules. Similar to MedLEE, MetaMap [2] is a rule-based system that extracts

* Corresponding author at: UAB 431, 1215 Western Ave, Albany, NY 12203, USA.
E-mail address: mtao@albany.edu (C. Tao).

http://dx.doi.org/10.1016/j.jbi.2017.07.002
1532-0464/

Table 1
Medication entries extracted from the prescription excerpts (1) and (2).

Medication name	Dosage	Mode	Frequency	Duration	Reason
Aspirin	325 mg	p.o.	4x/day	for 2 weeks	inflammation
Cefpodoxime	200 mg	through mouth	q.h.s.	a seven-day course	bronchitis

medical concepts (which includes medications) by querying the Unified Medical Language System (UMLS) Metathesaurus [3]. Both systems are unable to extract medication entries since they cannot interpret the relations of medication entities.

Research in automatic prescription extraction has been fostered by the Third i2b2 Challenge on NLP for Clinical Records [4]. The best performing system [5] in this challenge used a hybrid of machine learning classifiers and handwritten rules. It utilized Conditional Random Fields (CRFs) for medication information extraction and applied Support Vector Machines (SVMs) for relation extraction, reaching a horizontal phrase-level F1-measure of 0.857 on the i2b2 official benchmark set. Similarly, Li et al. [6] from the University of Wisconsin-Milwaukee trained CRFs with rules for medication information extraction but reached a relatively low performance (horizontal phrase-level F1-measure of 0.764). The significant performance differences using CRFs indicate the importance of system architecture, feature extraction, and parameter optimization. Besides, seven out of the top 10 systems, ranked from 2nd to 8th, were purely rule-based [7–13]. They utilized pattern matching rules with existing knowledge bases such as medication gazetteers.

2.2. Word embeddings in Named-Entity Recognition (NER)

Word embeddings [14] have been used in several NER tasks [15–17] to capture meaningful syntactic and semantic regularities using unsupervised learning from selected training corpora. In clinical NER, there are two prior studies that included word embeddings in their experiments.

First, De Vine et al. [18] analyzed the effectiveness of word embeddings in clinical concept extraction and studied the influence of various corpora used to generate embeddings. During feature extraction, they clustered word vectors into categories and used categorical labels as features for the classifier. They found that real-valued vectors did not show advantages when applied as feature set as the reason to use nominal categories via clustering. In our study, we want to evaluate the efficacy of real-valued word vectors instead of categorical labels, when directly used as classifier features. Second, Wu et al. [19] explored two neural word embedding algorithms (i.e., word2vec [20] and ranking-based [21]) in two clinical NER tasks. Both algorithms use a local context window model that do not explicitly consider global word-word co-occurrence statistics [22], which may contain important linguistic properties. In contrast, GloVe [23], introduced after word2-vec, specifically focused on the ratio of co-occurrence probabilities between different set of words when extracting word vectors. To the best of our knowledge, it is still unclear whether GloVe with real-valued vectors positively contribute to the clinical NER tasks.

3. Data

The Third i2b2 Challenge on NLP for Clinical Records [4] provided a corpus of 696 unannotated clinical records for development and 252 manually annotated records for testing. When participating into this challenge, Patrick et al. [5] manually annotated 145 developmental records and used them as training set. This training set contained 250,436 tokens, 21,077 medication entities, and 8516 medication entries. For testing, we used the official benchmark set from the i2b2 2009 challenge. See Table 2 for the per-category statistics.

4. Methods

We tackled prescription extraction in two consecutive steps: (1) medication information extraction, and (2) relation extraction. Fig. 1 depicts the workflow for our system. We present the details of each component below.

4.1. Pre-processing

We first pre-processed the data. We split the documents into sentences and then into tokens by simply splitting periods (excluding periods in acronyms, lists, and numbers) and whitespaces. We lowercased tokens and assigned part-of-speech (POS) tags using Natural Language Toolkit (NLTK) [24]. We replaced numbers, including literal and numerical forms, by placeholders (e.g., *five days → D days, 10am → DDam, 0.95 → .DD*).

4.2. Medication information extraction

We experimented with four different classifiers: Multinomial Naïve Bayes, SVMs, Decision Trees, and CRFs. We tuned the parameters of our classifiers using 5-fold cross validation on the training set. We then experimented with various feature sets and complemented our approach with post-processing rules. The results from each experiment were evaluated using phrase-level F1-measure (exact match, see section 4.4).

4.2.1. Feature extraction

Tokens and POS tags are the base features for our models. For durations and frequencies, we found that most phrases are introduced with and/or closed by specific signals. We captured this information by using two binary features representing whether the current token is a *starting* signal (e.g., *for, before, after*) or an *ending* signal (e.g., *weeks, days, hours*). We collected a list of these signals by harvesting the training data. Starting signals are mostly temporal prepositions, whereas ending signals tend to be names of time periods or clinical events (see Fig. 2 for additional examples).

Similarly, we extracted five more temporal binary features derived from the ones mostly used in the literature [25]. These features indicate whether a token represents a time (e.g., 8am, 7-pm), temporal period (e.g., decades, weekends), part of the day (e.g., morning, afternoon), temporal reference (e.g., today, yesterday), and numbers (e.g., 0.25, 700). Duration is one of the challenging medication categories. These signal features add characteristics to the tokens that belong to temporal expressions, which helps the classifier better identifying the beginning and the end of duration phrases.

Finally, we concluded feature extraction with the addition of word embeddings, which have been shown to capture meaningful syntactic and semantic regularities in NER tasks. In particular, we used GloVe to extract word vectors from MIMIC III [26]: a large critical care database, which among the others, contains about 2 million clinical notes for about 46 thousand patients. In contrast to other related studies, we pre-processed this dataset using the same normalizer applied on our own medication dataset to create