A Hybrid Approach to Extracting Disorder Mentions from Clinical Notes

Chunye Wang, PhD, Ramakrishna Akella, PhD
Department of Technology and Information Management
University of California Santa Cruz, Santa Cruz, CA 95064

Abstract

Crucial information on a patient’s physical or mental conditions is provided by mentions of disorders, such as disease, syndrome, injury, and abnormality. Identifying disorder mentions is one of the most significant steps in clinical text analysis. However, there are many surface forms of the same concept documented in clinical notes. Some are even recorded disjointedly, briefly, or intuitively. Such difficulties have challenged the information extraction systems that focus on identifying explicit mentions. In this study, we proposed a hybrid approach to disorder extraction, which leverages supervised machine learning, rule-based annotation, and an unsupervised NLP system. To identify different surface forms, we exploited rich features, especially the semantic, syntactic, and sequential features, for better capturing implicit relationships among words. We evaluated our method on the CLEF 2013 eHealth dataset. The experiments showed that our hybrid approach achieves a 0.776 F-score under strict evaluation standards, outperforming any participating systems in the Challenge.

Introduction

Clinical notes, such as discharge summary, radiology reports, echocardiogram reports, and electrocardiograph reports, are abundant in mentions of abnormalities, diseases, symptoms, and dysfunctions, which provide crucial disorder information on a patient’s physical or mental conditions. Identifying disorder mentions is one of the most significant steps in clinical text analysis. However, unlike the general newswire domain, where text is dominated by mentions of clearly defined locations, companies, and people names, there are many surface forms of the same concept documented in clinical notes. Some are even recorded disjointedly, briefly, or intuitively, due to the characteristics of natural language and the variety of written expression. For example, both “enlarged left atrium” and “left atrium … dilated” refer to the same concept “left atrial dilatation”; the disjointed mention “RV … enlarged” is written briefly and intuitively for the concept “right ventricular hypertrophy”. Such difficulties have challenged traditional information extraction systems that rely on explicit text features.

Background

There have been many natural language processing (NLP) systems developed to extract information from clinical text. cTAKES¹, MedEx², and MetaMap³ often use rule-based methods that rely on existing biomedical vocabularies to identify clinical entities. More recently, i2b2 and partners organized a few clinical NLP challenges to advocate the acquisition of increasingly fine-grained information from clinical records. The 2009 i2b2 NLP challenge aimed to extract information on medication of patients from discharge summaries, and further classify into seven fields such as medication names, doses, and frequencies. Among the top ten systems in the challenge, two were machine learning based systems, including the best one, while the second best used a rule-based method which extended the MedEx system. The 2010 i2b2 NLP challenge presented three tasks, where the concept extraction task focused on the extraction of patient medical problems, treatments, and tests. Gurulingappa et al trained conditional random fields (CRFs) on textual features enhanced with the output of a rule-based named entity recognition system⁴. deBruijn et al adopted a discriminative semi-Markov HMM to tag multitoken spans of text, as opposed to single tokens⁵. Some others utilized CRFs in an ensemble with existing NER systems, chunkers or different algorithms, with input based on knowledge-rich sources⁶⁻⁸. Later, on the same task, Jiang et al and Tang et al compared support vector machines (SVMs) and CRFs separately, and found both had similar performance⁹⁻¹⁰.

Although the 2010 i2b2 NLP challenge provided concept annotations to evaluate concept extraction task, these annotations were spread over different types of concepts, including medical problems, treatments, and tests. More importantly, concepts were defined as non-nested, non-overlapping, and contiguous words/phrases in clinical text, which is not always true in the real world, because there exist many variants of the same concept, as the disjointed examples above mentioned. Therefore, in our recent study, we focused on extracting all forms of mentions describing disorder concepts, especially with many being nested, overlapping, or disjointed. In this paper we present a hybrid approach to this task, which leverages supervised machine learning, annotation rules, and a standalone NLP system. Our contributions in this context are four-fold: (1) proposing new semantic features from ontology to
improve machine learning performance; (2) investigating the integration of a set of features, especially semantic, syntactic, and sequential features, for better catching implicit relationships among words; (3) developing a rule-based annotator which is automatically learned from training errors, and thus able to recognize patterns missed or over-learned by learning algorithms; and (4) overcoming the limitation of learning-based methods by employing an unsupervised NLP system for independent concept discovery from test data beyond the coverage of training data.

Methods

2013 ShARe/CLEF eHealth corpus

The 2013 ShARe/CLEF eHealth Task 1 corpus comprises of annotations over de-identified clinical reports from US intensive care units (version 2.5 of the MIMIC II database). The clinical notes consisted of discharge summaries, electrocardiogram, echocardiogram, and radiology reports. Each clinical note was annotated by two professional coders trained for this task, followed by an open adjudication step. A disorder mention was defined as the span(s) of text that can be mapped to a SNOMED-CT concept in UMLS and which belongs to the Disorder semantic group. The corpus of notes was split into 200 training and 100 test. Table 1 shows the distribution of annotations.

Table 1. Distribution of disorder mentions in training and test sets

|            | # Clinical Notes | # Total Mentions | # Unique Mentions* | # Disjointed Mentions |
|------------|------------------|------------------|-------------------|-----------------------|
| Training Data | 200              | 5721             | 2344              | 645                   |
| Test Data   | 100              | 5234             | 2055              | 432                   |

* Unique mentions are those only appear in either training or test data set.

System framework

We present the framework of our hybrid extraction system and data flow in Figure 1. The system is mainly composed of three extraction components and a post-processing module. Firstly, SVMs are learned from training data to predict if a word belongs to disorder concepts. Seven types of features (explained later) are extracted in order to help SVMs acquire both explicit and implicit relationships among words. Then, a rule-based annotator is automatically constructed from frequent errors SVMs made on training data for error correction. Finally, we employ an unsupervised NLP system, MetaMap, as the supplement, to discover concepts unique in test data. The outputs from above three components will be compiled in the post-processing module to determine mention boundaries.

Supervised machine learning

We formulated the extraction task as a sequence labeling problem which can be defined as follows: given a sequence of input words \( x = (x_1, ..., x_n) \), and a set of labels \( L \), determine a sequence of labels \( y = (y_1, ..., y_n) \), for the words such that \( y_i \in L \) for \( 1 \leq i \leq n \). The label \( y_i \) here contains two pieces of information: the type of the label and the position of the word in the sequence. BIO label formalism is widely used to transform mention annotations to a machine-readable format\(^5\)\(^9\)\(^11\). If a word is the Beginning word/Inside/Outside of a mention, it gets the label B, I, or O. However, when predicting disjointed mentions, which are frequent in our study, BIO is prone to determining inaccurate mention boundaries, because the end of a disjointed mention not only depends on I, but also B of the next mention. To solve this problem, another two labels E and S are added to represent the End word of a mention and
the Single word mention, so that the end of a disjointed mention will only depend on E, and the single world mention will only depend on S. Then the extraction task becomes a classification task to assign each word with one of the five labels \( L = \{B, I, E, S, O\} \) based on its characteristics and context. Support vector machines are chosen for classification due to its good performance on clinical information extraction\(^{10,11}\).

### Integration of rich features

We generated various types of features and systematically investigated their gains on concept extraction. Basically, the explicit features, such as Bag-of-Words (BOW), orthographic and morphologic features, are from word itself, but unable to provide semantics, syntax, or sequential information. Hence we further extracted implicit features from context, such as Part-of-Speech (POS), sequential, and semantic features. Table 2 describes every type of feature.

#### Semantic features from ontology

We investigated ontologies to bridge the gap between machine learning and semantic understanding. We used SNOMED-CT to generate two types of semantic features, the semantic type features (ST) and the semantically related term features (SRT), to improve the performance of learning algorithms. Particularly, we proposed the novel semantically related term features to expand the coverage of training annotations. At first, a disorder concept from training annotations is mapped to a concept node in the ontology; words describing its parents and/or children nodes are obtained as the semantically related words of the concept; when determining the feature value of a target word, we check whether it belongs to the semantically related words of the concept. If so, this word will be associated with that disorder concept by such a feature. Thus, even if the word itself does not appear in any training annotations, learning algorithms are able to capture its connection to a concept through SRT features.

| Feature Type                          | Description                                                                 | Examples                                                                 |
|---------------------------------------|-----------------------------------------------------------------------------|-------------------------------------------------------------------------|
| BOW                                   | words of disorder mentions in training set                                  | “dizziness”; “edema”; “facial”                                          |
| Orthographic features                 | whether a word contains capital letters, digits, special characters, etc.    | contain digit (“s3”), initial capital (“B-cell”), all capital (“MR”),    |
|                                       |                                                                             | contain hyphen (“T-wave”), CapsAndDigits (“DM2”)                         |
| Morphologic features                  | whether a word contains certain prefix or suffix                           | contain anti- (“antigen”, “anticoagulation”, “anti-inflammatory”)        |
|                                       |                                                                             | contain -ous (“granulomatous”, “edematous”, “erythematous”)              |
| Part-of-Speech                        | Part-of-Speech tag of a word                                               | “epigastric ventral hernia” [JJ JJ NN]                                  |
| Sequential Features                  | label assigned to previous word                                            | B, I, E, S, or O                                                        |
| Semantic Type Features                | semantic categories of words (defined in ontology)                         | “pericardial effusion” [Disease or Syndrome]; “allergies” [Pathologic Function] |
| Semantically Related Term Features    | whether a word belongs to semantically related words of a disorder concept in training data (obtained from parents and/or children nodes in ontology) | whether a word is one of [{“cerebral”, “degeneration”, “dementia”, “senile”, “presenile”, “aphasia”}], which are semantically related words of “Alzheimer”, a disorder concept in training data |

#### Rule-based annotator

SVMs are capable of identifying concept mentions, but still make a number of false positive and false negative errors. After error analysis, we found that part of these errors could be corrected subsequently using annotation rules. We grouped misclassified mentions, counted their frequencies, automatically generated regular expression rules\(^4\) from frequent error mentions, and assigned true labels to the text spans that match the rules. To avoid overfitting, after applying a rule, we inspected the ratio of the number of true annotations given by the rule over the number of false annotations given by the rule, and set a threshold to control whether this rule should be included.

#### MetaMap

Because both SVM and rule-based annotator are dependent on training data, their generalization capability is theoretically limited when applied to test data. Consequently, we added MetaMap as the third extraction component for independent concept discovery from the text beyond the coverage of training data. MetaMap is an NLP system developed to link the text of medical documents to the knowledge embedded in UMLS Metathesaurus. It employs knowledge bases, NLP, and computational-linguistic techniques, and automatically discovers concepts after shallow parsing, NP chunking, and lexicon lookup, in an unsupervised manner.

### Results

We evaluated our hybrid extraction system on 2013 CLEF eHealth data set. SVM with RBF kernel was learned from training data using LIBSVM library. The cost and Gamma parameters were tuned under the five-fold cross validation, and other parameters were set to default values. The test data set was withheld for evaluation purpose.
only. We conducted a series of experiments to validate the proposed features and system components. The outputs of the disorder mention extraction system were compared with the gold standard annotations by strict and relaxed matching standards. The strict matching requires both the extracted mention and the gold standard have identical start and end words, while the relaxed matching only requires they have an overlap. Performance was evaluated in terms of precision, recall, and F1 score. Since our system was tuned and optimized for F1 score under the strict matching standard, we will discuss and compare in this context below, unless otherwise stated. Table 3 summarizes the experimental results of different system settings on test data. SVM using only Bag-of-Words features is set as baseline and F1 improvement over it ( ΔF1 ) is calculated for comparison convenience.

### Performance of SVMs with different features

The baseline setting achieved a 0.444 F1 score, compared with which each additional type of features improved the performance. After the comparison of individual feature types, the gains from implicit features (POS, sequential, SRT, or ST) were consistently higher than those from explicit features (orthographic or morphologic). Especially, semantic type features significantly increased F1 score by 31%. The proposed semantically related term features also proved effective and better than traditional orthographic and morphologic features. The learning algorithm integrating all types of features finally achieved a 0.644 F1 score, 45% improvement over the baseline.

### Performance of additional extraction components

The rule-based annotator was constructed upon prediction errors on training data. The evaluation on test data showed 0.146 and 0.120 increases over the baseline and SVM using all features, respectively, which made it a powerful supplement to the system. In addition, the training-independent component MetaMap further lifted F1 by 0.200 (45%) improvement over the baseline.

### Discussion

Concept recognition is a fundamental step in clinical text analysis. With the prevalence of annotated records, more and more clinical information extraction systems begin to adopt machine learning approach. However, in real-world application, we should be aware of the limitations of learning-based methods. Specifically, when data is sparse, trained models tend to lose generalization capability on test data; when the entities have many surface forms, such as concept mentions recorded disjointedly, briefly, or intuitively, traditional features in learning algorithms become insufficient to capture such variations. In our work, we attempted to overcome these challenges as explained below.

On the system level, we proposed two additional extraction components to supplement machine learning methods. (1) SVM with all features could achieve a 0.644 F1 score, which appears reasonably good, but is actually very good when considering the fact that 41% training annotations do not appear in test annotations, and 39% test annotations do not appear in training annotations (see Table 1), which makes the task very challenging. However, there were still a number of false positive and false negative errors due to the sparse data leading to missed or over-learnt features. The rule-based annotator, which was constructed from frequent errors in an unsupervised manner, had a 0.834 precision, a 0.199 recall, and a 0.321 F1 score on mislabeled data. Incorporating rule-based annotator, the system performance increased to 0.764 instantly. (2) To make the system generalize better on new data, we employed

### Table 3. Performance of extraction system with different feature and component settings

| SVM Features | Rules | Meta Map | P/R/F1 (strict) | ΔF1 (%) | P/R/F1 (relaxed) |
|--------------|-------|----------|----------------|---------|-----------------|
|             |       |          |                |         |                 |
| +            | +     |          | 0.463/0.476/0.469 | 0.025 (6%) | 0.817/0.813/0.815 |
| +            |       |          | 0.465/0.473/0.469 | 0.025 (6%) | 0.814/0.806/0.810 |
| +            | +     |          | 0.480/0.507/0.493 | 0.049 (11%) | 0.799/0.816/0.807 |
| +            | +     | +        | 0.500/0.486/0.493 | 0.049 (11%) | 0.859/0.823/0.841 |
| +            | +     | +        | 0.466/0.481/0.473 | 0.029 (7%) | 0.809/0.813/0.811 |
| +            | +     | +        | 0.611/0.556/0.582 | 0.138 (31%) | 0.876/0.790/0.831 |
| +            | +     | +        | 0.571/0.611/0.590 | 0.146 (33%) | 0.825/0.865/0.845 |
| +            | +     | +        | 0.506/0.531/0.518 | 0.074 (17%) | 0.815/0.834/0.824 |
| +            | +     | +        | 0.463/0.472/0.467 | 0.023 (5%) | 0.820/0.810/0.815 |
| +            | +     | +        | 0.636/0.575/0.604 | 0.160 (36%) | 0.887/0.791/0.836 |
| +            | +     | +        | 0.698/0.598/0.644 | 0.200 (45%) | 0.920/0.757/0.831 |
| +            | +     | +        | 0.816/0.719/0.764 | 0.320 (72%) | 0.935/0.825/0.877 |
| +            | +     | +        | 0.816/0.740/0.776 | 0.332 (75%) | 0.929/0.844/0.884 |

Orth. - orthographic feature; Mor. - morphologic feature; Seq. - sequential feature.
MetaMap as a standalone NLP system to identify disorder mentions not shown in training data. In experiments on unique mentions in test data, MetaMap attained a 0.801 precision, a 0.183 recall, and a 0.298 F1 score. The low recall is mainly caused by the inaccurate NP chunking and the inability to recognize and label disjointed phrases. With the assistance of MetaMap, the final integrated system achieved a 0.776 F1 score under strict matching standard (0.764 if excluding external annotations from MetaMap), outperforming any participating systems in the 2013 ShARe/CLEF eHealth Challenge (0.750 the best result from University of Texas Health Science Center)\textsuperscript{12}. The encouraging results demonstrate the effectiveness of our synergistic extraction framework.

On the learning algorithm level, we integrated various types of features for better identifying concept variations. The baseline model had an F1 score of 0.799 by relaxed matching, but only 0.444 by strict matching. This indicates that often only a part of the mention can be identified. It is difficult to obtain all text spans of a concept mention using bag-of-words features only. Adding extra explicit features helped, but very little (ΔF1=0.023). Therefore, explicit features are insufficient to capture the underlying relationships among the spans of concept variants. We sought to characterize these relationships from multiple perspectives, including semantics, syntax, and sequence. The experimental results substantiated that the implicit features from these facets were very beneficial (ΔF1=0.160).

To advance the semantic understanding of clinical text, we explored ontological resources (e.g. SNOMED-CT) to generate two types of semantic features. In SRT features, semantically related words of a concept were acquired from its parent and children nodes in ontology, which were used to expand the “span” of training annotations. In ST features, the semantic type of a concept was obtained from ontology as a feature. In experiments, these two types of features significantly boosted F1 by 34%, contributing more than any other features or extraction components. This confirms the necessity of semantic understanding in mining clinical text, and validates our approach to utilizing ontology. We believe that the abundant ontological resources have significant potential to exploit, and our study on them is still at an early stage. In future, we would like to target more specific and relevant ontology to the task.

Conclusion

In this paper, we investigated approaches to extracting disorder mentions from clinical notes, and systematically evaluated the effects of different features and extraction components. Our final solution integrated supervised machine learning, annotation rules, and an unsupervised NLP system, in a synergistic manner. The implicit features we generated, especially those from ontological resources, are of great assistance in capturing underlying relationships among words to recognize disjointed mentions. The system produced a state-of-the-art performance with a 0.776 F1 score as evaluated in 2013 ShARe/CLEF eHealth Challenge.

References

1. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-13.
2. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc 2010;17(1):19-24.
3. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229-36.
4. Wang C, Akella R. UCSC's System for CLEF eHealth 2013 Task 1. Labs Working Notes of Conference and Labs of the Evaluation Forum 2013.
5. Gurulingappa H, Hofmann-Apitius M, Fluck J. Concept identification and assertion classification in patient health records. Proc i2b2/VA Workshop on Challenges in NLP for Clinical Data 2010.
6. deBruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. J Am Med Inform Assoc 2011;18(5):557-62.
7. Jiang M, Chen Y, et al. Hybrid approaches to concept extraction and assertion classification - vanderbilt’s systems for 2010 i2b2 NLP Challenge. Proc i2b2/VA Workshop on Challenges in NLP for Clinical Data 2010.
8. Torri M, Liu H. BioTagger-GM for detecting clinical concepts in electronic medical reports. Proc i2b2/VA Workshop on Challenges in NLP for Clinical Data 2010.
9. Jiang M, Chen Y, Liu M, et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. J Am Med Inform Assoc 2011;18(5):601-6.
10. Tang B, Can H, et al. Clinical Entity Recognition using Structural Support Vector Machines with Rich Features. Proc of the ACM sixth international workshop on data and text mining in biomedical informatics 2012:13-20.
11. Doan S, Collier N, Xu H, Duy PH, Phuong TM. Recognition of medication information from discharge summaries using ensembles of classifiers. BMC Med Inform Decis Mak 2012;12:36.
12. Pradhan S, Elhadad N, South B, et al. Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. Proceedings of ShARe/CLEF eHealth Evaluation Labs 2013.