A Hybrid System for Named-Entity Extraction from Clinical Discharge Summaries

Meliha Yetisgen-Yildiz, PhD[1], Safiyyah Saleeem, MSc[2], Daniel Capurro, MD[1]
[1]Biomedical & Health Informatics, School of Medicine, University of Washington, Seattle, WA
[2]Department of Linguistics, University of Washington, Seattle, WA

Abstract

Identification of medical entities is a critical step to extract the information represented in the clinical text. In this paper, we describe a statistical system that identifies three core medical entity types, medical problem, treatment, and medication, in clinical discharge summaries. We developed our system and evaluated its performance with the annotated corpus created for this year's i2b2 Natural Language Processing (NLP) challenge. We increased the performance of the statistical system with heuristics based on the knowledge available in Unified Medical Language System (UMLS).

Introduction

Within the last decade, Electronic Medical Records (EMR) are becoming integral components of health care services. The American Medical Association recently published a white paper with key recommendations about secondary uses of EMR[1]. Analysis of health care information available in EMR systems can expand knowledge about diseases and potential treatment, improve understanding of effectiveness and efficiency, and support public health and safety goals[2]. To accomplish those recommendations, accessibility to patient data available in EMR systems plays a critical role. However most patient information is represented in free-text form and textual information is difficult for automated approaches to reliably access. A representation step is required to convert the unstructured information available in free-text into a structured format so that the automated approaches can work on. In this paper, we present a hybrid named-entity recognition system that identifies medical entities in text. Our system uses statistical approaches to identify the spans that represent the entities and post process the identified spans with heuristics based on a medical knowledge base, UMLS[3]. We used 2010 i2b2 Natural Language Processing (NLP) Challenge corpus to train and evaluate our system for three medical entity types; *medical problem*, *treatment*, and *test*. The i2b2 corpus is composed of annotated clinical discharge summaries from four different hospitals. In the following sections, we will describe the details of our system and present its performance.

Named Entity Recognition

Named entity recognition (NER) is a well-known task of information extraction. Traditionally, the task has been based on identifying words and phrases that refer to various entities of interest, including persons, locations, and organizations[4]. Previous works have tackled the NER within the biomedical domain[5], clinical domain[6], newswire domain[7], and email domain[8]. Much of the NER work in the biomedical domain has been done on identifying genes and proteins in the scientific literature.

As the benefit of structuring data in EMR systems has become more evident, work has increased on NER from unstructured clinical notes and discharge summaries generated by clinicians. Clinical language and vocabulary is highly complex and rapidly evolving, making the identification of entities a difficult task.

The ability to recognize previously known entities is an essential part of NER. Such ability is based upon the recognition and classification of rules triggered by distinctive features associated with positive and negative examples. Early NER research in the clinical domain is mainly based on manually generated rules. Although, rule-based recognition systems produce extraction results with high precision in the domains they are designed for, they perform poorly in other domains.

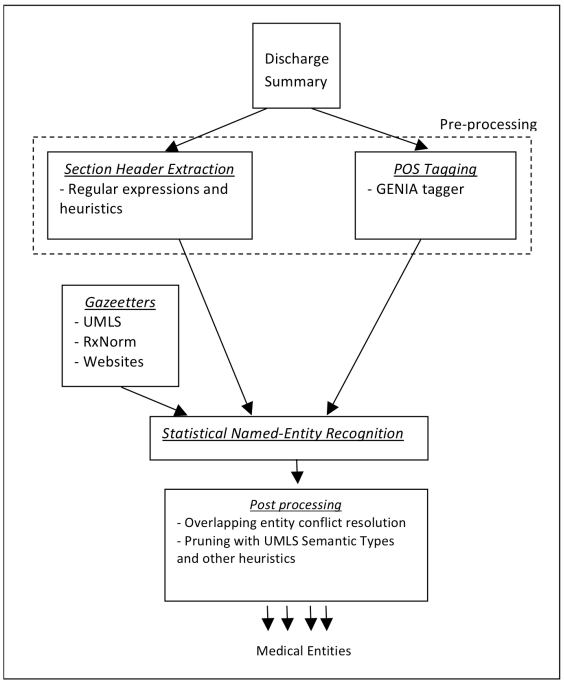
As an alternative approach, supervised machine learning approaches have been widely used to identify named-entities in text. However, supervised approaches depend crucially on the availability of properly annotated corpora. In the clinical domain, this means that large collections of online patient records in textual form must be available to human annotators. However, patient records are confidential, and in order to make them accessible for research purposes personal information must be removed to comply with the laws protecting patient confidentiality. The i2b2 NLP challenges play an important role in increasing the momentum of

clinical NLP field by providing annotated corpora to researchers[9-11].

Method

We built a hybrid system that combines rule-based approaches with statistical NER approaches to identify medical entities. The overall system architecture is presented in Figure 1. Our system first pre-process discharge summaries to identify the main section headings (i.e. Past Medical History, Meds on Admission) to chunk the reports into sections. It also uses the Genia Tagger[12] to identify the POS tags of all the tokens in text. In the next step, it uses the pre-processing output and gazetteers as features in a statistical NER approach to identify the spans that represent *medical problems*, *treatments*, and *tests* in text. Finally, it post-processes the identified spans to filter the false positives and to correct the boundaries of the spans.

Figure 1. Overall system architecture for named entity extraction.



Section Header Extraction

To identify the characteristics of the section headers, the third author of the paper manually went over a subset of the discharge summaries available in the i2b2 training set and annotated the section headings (i.e. *Chief Complaint, Past Medical/Surgical History)*. He identified 14 general categories from his annotations and mapped each section heading to one of those 14 general categories (i.e. Category *Past Medical/Surgical History* includes *allergies, Cardiac Risk Factors, rf, past surgical history, past obstetric history, past gyn history*). Based on this knowledge, we designed regular expressions to detect the characteristics of the headers (i.e. capitalization patterns) and used the common keywords for each general category to generate rules that can map each identified section header to a category.

Gazetteer Construction

The dictionary features define various categories of words including, *medical problems*, *treatments*, *tests*, and *drug names*. We constructed those lists by querying UMLS[3] and RxNorm[13]. We used the semantic types listed in the i2b2 concept annotation guidelines to select the UMLS terms related to each entity type (Table 1).

Entity Type	Semantic Types	Dictionary
Medical Problem	- Acquired Abnormality - Anatomical Abnormality - Cell or Molecular Dysfunction - Congenital Abnormality - Disease or Syndrome - Injury or Poisoning - Mental or Behavioral Dysfunction - Neoplastic Process - Pathologic Function - Sign or Symptom - Virus	MeSH SNOMED ICD9
Treatment	- Antibiotic - Biomedical or Dental Material - Clinical Drug - Drug Delivery Device - Medical Device - Pharmacologic Substance - Steroid - Therapeutic or Preventive Procedure	MeSH SNOMED RxNorm ICD9 ICD10 NCI
Test	- Diagnostic Procedure - Laboratory Procedure	MeSH SNOMED ICD9

Table 1. Semantic types and dictionaries used in gazetteer construction.

To prevent noise in our lists, we restricted our search for the dictionaries listed in Table 1 and selected only preferred names of UMLS concepts. Regular expressions tailored to the format of each dictionary were used to process the gazetteers by removing simple characters such as parentheses, commas, brackets, etc; remove certain modifiers such as *accidental*, *other*, *unspecified*, etc; and split apart

entries that contained conjunctions (i.e. *and*, *or*) to appear as separate items. For example, after cleaning, the SNOMED concept *(Abscess) or (boil) or (furuncle) of eyelid* would be *abscess*, *boil*, *furuncle of eyelid* each of which appeared as separate items in the gazetteer.

In addition to medical gazetteers extracted from UMLS, we created other lists of gazetteers to identify the boundaries of the entity spans (i.e. determiners, possessive pronouns, conjunctions, measurement abbreviations).

Named Entity Recognition

We posed the problem of medical NER, as a sequence labeling task similar to the part-of-speech (POS) tagging or phrase chunking tasks, where each token in the input text corresponds to a label in the output, and is solved with sequential classification algorithms. We picked Conditional Random Fields (CRF) [14] as the sequential classifier in our NER task. CRFs are undirected graphical models used to calculate the conditional probability of values on designated output nodes given values assigned to other designated input nodes. In the special case in which the output nodes of the graphical model are linked by edges in a linear chain, CRFs make a first order Markov independence assumption, and thus can be understood as conditionally trained finite state machines. In our system, we used the CRF implementation available in MinorThird[15].

Learning with CRF reduces NER to the task of tagging each token in a given document. We used a set of three tags to define the spans in our classification, corresponding to (1) a begin token, (2) a continue token, (3) a token that is not part of an entity. The set of features used in training NER models are presented in Table 2. All features were instantiated for each token *t*, as well as for a window of three tokens to the left and three tokens to the right of token *t*.

The basic features included the lower/upper-case value of a token *t*, its capitalization pattern, constructed by replacing all capital letters with the letter *X*, all lower-case letters with the letter *x*, all digits with *9*, and its length. The gazetteer features were constructed from the gazetteers described in the previous section. We also defined report features to capture the characteristics of the numeric tokens and header tokens.

Basic Features
lexical value, lowercase – eg. f(t=cancer)=1 lexical value, uppercase – eg. f(*t*=EKG) = 1 capitalization pattern – eg. t=Aspirin, f(t.cap=Xx+)=1 character length
Gazetteer Features
t in MedicalProblemUMLSList – eg. *migraine* *t* in TreatmentUMLSList – eg. *surgery* *t* in TestUMLSList – eg. *X-ray* *t* in DrugListRxNorm – eg. *Tylenol* *t* in BoundaryTokenList – eg. *the* (determiner), *and* (conjunction), *her* (possessive pronoun) *t* in MeasurementAbbreviationList – eg. *mg*
Report Features
t appears in section heading *t* is an acronym *t* is a decimal number *t* is a real number *t* is a percentage number *t* represents an interval

Table 2. Feature sets.

Extraction Errors and Post-Processing Heuristics

While building the system, we used the 2010 i2b2 NLP challenge annotated training corpus. We first divided the annotated corpus into training (90%) and test (10%) sets. We trained the extractor models by using the training set and analyzed its extraction errors by using the test set. After this investigation, we identified the following classes of errors and proposed post-processing solutions based on heuristics to fix them.

1. False Positives:
   a. Prediction errors: While investigating the false positive prediction errors, we observed that the majority of those errors could be solved with the knowledge available in the UMLS Semantic Network[16]. For example if an identified span has the semantic type *Clinical Drug*, it should only be identified as a *treatment*, not as a *medical problem* or a *test*.

      We first created semantic type lists associated with each entity type by querying UMLS for the gold standard spans present in the annotated training corpus. Querying UMLS for semantic types was not a trivial task since most of the spans started with determiners, possessive pronouns, or pronouns due to the i2b2 annotation guidelines used while annotating the corpus. We pruned those head tokens of the spans hence they were not part of UMLS concepts and queried UMLS to extract the associated semantic types.

Exp.	Medical Problem						Treatment						Test					
	Exact			Inexact			Exact			Inexact			Exact			Inexact		
	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
1	**0.76**	0.79	0.77	**0.87**	0.90	0.88	**0.78**	0.79	0.78	**0.87**	0.89	0.88	**0.77**	0.83	**0.80**	0.85	0.90	0.87
2	0.75	**0.82**	0.78	0.86	**0.94**	**0.90**	0.77	**0.82**	**0.80**	0.86	**0.93**	**0.90**	0.75	**0.86**	**0.80**	**0.86**	0.93	**0.90**
3	0.75	**0.82**	**0.79**	0.86	**0.94**	**0.90**	0.74	**0.82**	0.78	0.84	**0.93**	0.88	0.75	**0.86**	**0.80**	0.83	**0.95**	0.88

Table 3. Performance evaluation. P: Precision, R: Recall, F: F-measure; the highest value for each column is in boldface (Exp #1: Model trained with training set without post processing, Exp #2: Model trained with training set with post processing, Exp #4:Model trained with merged set with post processing)

For each entity type, our system extracted the semantic types of the identified spans from UMLS, compared the extracted semantic types against the semantic type list of the entity type and pruned the ones with no matching semantic types.

b. Spans identified for multiple entity types: There were many cases where the same span was identified for multiple entity types. As an example, the medical concept, *barbiturates,* was identified both as a medical problem (true positive for medical problem) and as a treatment (false positive for treatment). Semantic type filtering eliminated the majority of such errors. For the remaining ones, we checked the prediction probabilities and picked the entity type with the highest prediction probability as the true entity type and pruned the other ones.

2. Exact Span Errors

a. Error in the first token of the span (i.e. ***is*** *mildly dilated*, *+ elevated JVD*, ***of*** *prostate cancer*): We checked the first token of the each identified span and pruned its first token if it is a modal verb, preposition, or non-letter character.

b. Span that includes multiple entities merged by conjunctions (i.e. *a left non displaced sacral wing fracture* **and** *left inferior pubic ramus fx*): We solved this type of errors by implementing a post-processing step that searches a span for conjunctions (i.e. *and*) and chunks the spans into sub-spans based on the location of the conjunctions.

Evaluation

This year's i2b2 corpus was composed of 349 annotated training documents, 477 annotated test documents, and 827 unannotated documents. We built the NER models by using the training and unannotated sets and tested the performance of the models with the test set. We used precision, recall and F-1 measures to calculate the overall performance for both exact and inexact spans. Table 3 presents the performance values for three different experiment configurations. For all entity types, the performance values for inexact spans were higher than those of exact spans.

In our experiments, we first measured the effect of our post-processing heuristics. To accomplish this, we compared the performance values of our statistical system without (exp #1) and with (exp #2) post-processing heuristics. As can be seen from Table 3, using post-processing heuristics had a positive effect (~3%) on both exact and inexact precision of NER for all entity types.

In our third experiment, we tried to extent the training set with the unannotated set. We first created NER models by using the training set and used those models to annotate the unannotated set. Next, we manually corrected the false positives in the annotations and merged this set with the training set. We used this merged set to create new NER models and measured their performance (exp #3). Because we manually corrected only false positives, we didn't receive positive performance changes. The precision and recall didn't change for medical problem, and slightly decreased for treatment. We plan to revise our correction approach and repeat the same experiment as future work.

Conclusion

Extracting entities in free-form text is the most critical step in converting the unstructured representation of knowledge available in text into a computable representation. In this paper, we presented our hybrid system that combines statistical approaches with heuristics to identify medical named entities in clinical discharge summaries. We achieved good performance results for inexact spans and plan to refine our approach to improve the performance results for the exact spans.

References

1. Safran C., Bloomrosen M, Hammond WE, et. al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. J Am Med Inform Assoc. 2007; 14(1): pp. 1-9.
2. Herasevich V, Pickering BW, Dong Y, Peters SG, Gajic O. Informatics Infrastructure for Syndrome Surveillance, Decision Support, Reporting, and Modeling for Critical Illness. Mayo Clinic Proc. 2010; 85(3): pp. 247-254.
3. UMLS Fact Sheet. National Library of Medicine. 2006. Available at: http://www.nlm.nih.gov/pubs/factsheets/umls.html.
4. Nadeau D, Sekine S. A survey of named entity recognition and classification. Linguisticae Investigationes 2007; 30(1): pp. 3-26.
5. Settles B. Biomedical named entity recognition using conditional random fields and rich feature sets. Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, Geneva, Switzerland, August 2004.
6. Halgrim S, Xia F, Solti I, Cadag E and Uzuner O. Extracting Medication Information from Discharge Summaries. Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents, Los Angeles, June 2010. pp. 61-67.
7. Grishman R and Sundheim B. Message Understanding Conference-6: a brief history. In Proceedings of the 16th Conference on Computational Linguistics. Morristown, NJ, 1996. pp. 466-471.
8. Minkov E, Wang RC, and Cohen WW. Extracting personal names from email: applying named entity recognition to informal text. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Vancouver, October 2005. pp. 443-450.
9. Uzuner O, Goldstein I, Lua Y, Kohane I. Identifying patient smoking status from medical discharge records. Journal of the American Medical Informatics Association. 2008; 15(1): pp. 14-24.
10. Uzuner O. Recognizing obesity and comorbidities in sparse data. Journal of the American Medical Informatics Association. 2009; 16(4): pp. 561-570.
11. Uzuner O, Solti I, Cadag E. i2b2 NLP Challenges. Third i2b2 Workshop: Challenges in Natural Language Processing for Clinical Data Medication Extraction Challenge. San Francisco, 2009.
12. Tsuruoka Y, Tateishi Y, Kim JD, Ohta T, McNaught J, Ananiadou S, Tsujii J. Developing a Robust Part-of-Speech Tagger for Biomedical Text. Advances in Informatics - 10th Panhellenic Conference on Informatics, 2005, LNCS 3746, pp. 382-392.
13. RxNorm. National Library of Medicine. 2010. Available at: http://www.nlm.nih.gov/research/umls/rxnorm/
14. Lafferty J, McCallum A. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of ICML, 2001.
15. William C. Minorthird: Methods for Identifying Names and Ontological Relations in Text using Heuristics for Inducing Regularities from Data, 2004. Available at: http:// minorthird.sourceforge.net/.
16. UMLS Semantic Network Fact Sheet. National Library of Medicine. 2006. Available at: http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html