ezDI: A Hybrid CRF and SVM based Model for Detecting and Encoding Disorder Mentions in Clinical Notes

Parth Pathak, Pinal Patel, Vishal Panchal, Narayan Choudhary, Amrish Patel, Gautam Joshi
ezDI, LLC.
{parth.p,pinal.p,vishal.p,narayan.c,amrish.p,gautam.j}@ezdi.us

Abstract

This paper describes the system used in Task-7 (Analysis of Clinical Text) of SemEval-2014 for detecting disorder mentions and associating them with their related CUI of UMLS[1]. For Task-A, a CRF based sequencing algorithm was used to find different medical entities and a binary SVM classifier was used to find relationship between entities. For Task-B, a dictionary look-up algorithm on a customized UMLS-2012 dictionary was used to find relative CUI for a given disorder mention. The system achieved F-score of 0.714 for Task A & accuracy of 0.599 for Task B when trained only on training data set, and it achieved F-score of 0.755 for Task A & accuracy of 0.646 for Task B when trained on both training as well as development data set. Our system was placed 3rd for both task A and B.

1 Introduction

A clinical document contains plethora of information regarding patient's medical condition in unstructured format. So a sophisticated NLP system built specifically for clinical domain can be very useful in many different clinical applications. In recent years, clinical NLP has gained a lot of significance in research community because it contains challenging tasks such as medical entity recognition, abbreviation disambiguation, inter-conceptual relationship detection, anaphora resolution, and text summarization. Clinical NLP has also gained a significant attraction among the health care industry because it promises to deliver applications like computer assisted coding, automated data abstraction, core/quality measure monitoring, fraud detection, revenue loss prevention system, clinical document improvement system and so on.

Task-7 of SemEval-2014 was in continuation of the 2013 ShaRe/CLEF Task-1 (Sameer Pradhan, et al., 2013). This task was about finding disorder mentions from the clinical text and associating them with their related CUIs (concept unique identifiers) as given in the UMLS (Unified Medical Language System). UMLS is the largest available medical knowledge resource. It contains 2,885,877 different CUIs having 6,497,937 different medical terms from over 100 different medical vocabularies. Finding accurate CUIs from free clinical text can be very helpful in many healthcare applications. Our aim for participating in this task was to explore new techniques of finding CUIs from clinical document.

Over the last few years many different Clinical NLP systems like cTAKES (Savova, Guergana K., et al., 2010), MetaMap (A. Aronson, 2001), MedLEE (C. Friedman et al., 1994) have been developed to extract medical concepts from a clinical document. Most of these systems focus on rule based, medical knowledge driven dictionary look-up approaches. In very recent past, a few attempts have been made to use supervised or semi-supervised learning models. In 2009, Yefang Wang (Wang et al., 2009) used cascading classifiers on manually annotated data which fetched F-score of 0.832. In 2010, i2b2 shared task challenge focused on finding test, treatment and problem mentions from clinical document.

In 2013, ShARe/CLEF task focused on finding disorder mentions from clinical document and assigning relevant CUI code to it. In both i2b2 task and ShaRe/CLEF task most of the systems used either supervised or semi-supervised learning ap-

[1] http://www.nlm.nih.gov/research/umls/

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 278–283,
Dublin, Ireland, August 23-24, 2014.

proaches.

In this paper we have proposed a hybrid supervised learning approach based on CRF and SVM to find out disorder mentions from clinical documents and a dictionary look-up approach on a customized UMLS meta-thesaurus to find corresponding CUI.

2 Data

The SemEval-2014 corpus comprises of de-identified plain text from MIMIC[2] version 2.5 database. A disorder mention was defined as any span of text which can be mapped to a concept in UMLS and which belongs to the Disorder semantic group. There were 431 notes extracted from intensive care unit having various clinical report types (like radiology, discharge summary, echocardiogram and ECG), out of which 99 notes were used in development data set, 199 notes were used in training data set and 133 notes were used in testing data set.

Preliminary analysis on this data showed that number of sentences in training documents were comparatively smaller than the development or test data set (Table 1). Number of disorder mentions were also significantly lower in training data set than in development data set (Table 1).

Type	Dev	Train	Test
Docuemnts	99	199	133
Sentence	9860	10485	17368
Token	102k	113k	177k
Avg token/sen	10.42	10.79	10.24
Cont. entity	4912	5,165	7,186
Disjoint Entity	439	651	4588
Avg Ent/Doc	54.05	29.22	57.47
Distinct CUI	1007	938	NA

Table 1: Numerical analysis on data.

3 System Design

Analysis of Task-A showed that disorder mentions also contain other UMLS semantic types like findings, anatomical sites and modifiers (Table 2). So we divided the task of finding disorder mention in to two subtasks. First a CRF based sequencing model was used to find different disorder mentions, modifiers, anatomical sites and findings. Then a binary SVM classifier was used to check if relationship exists between a disorder and other types of entities or not.

Example	Disorder	Findings	Anatomy	Modifier
There is persistent **left lower lobe opacity** presumably **atelectasis**.	✓	✓	✗	✗
He had substernal **chest pain**, sharp but without **radiation**.	✓	✓	✗	✗
Patientt also developed some **erythema** around the **stoma** site on hospital day two.	✓	✗	✓	✗
The tricuspid valve **leaflets** are mildly **thickened**.	✗	✓	✓	✗
Please call,if you find **swelling** in the **wound**.	✓	✓	✗	✗
She also notes new sharp **pain** in **left shoulder blade**/back area.	✓	✗	✓	✗
An echocardiogram demonstrated mild **left** and **right atrial dilatation**	✓	✗	✗	✓

Table 2: Entity Types co-relation and examples

For Task-B, we have used a simple dictionary look up algorithm on a customized UMLS dictionary. A preliminary analysis of UMLS entities in general show that a single disorder mention may consist of various types of linguistic phrases. It is not necesarry that the system to detect these entities as a single phrase. The entities and their relations may also occur in disjoint phrases as well. Our analysis of the disorder entities inside UMLS reveals that out of a total 278,859 disorders (based on SNOMED-CT library), 96,069 are such that can be broken down into more than one phrase, which is roughly 1/3 of total number of disorders in the UMLS.

3.1 System Workflow

The Work-flow of the system is as follow:

[2] http://mimic.physionet.org/database/releases/70-version-25.html

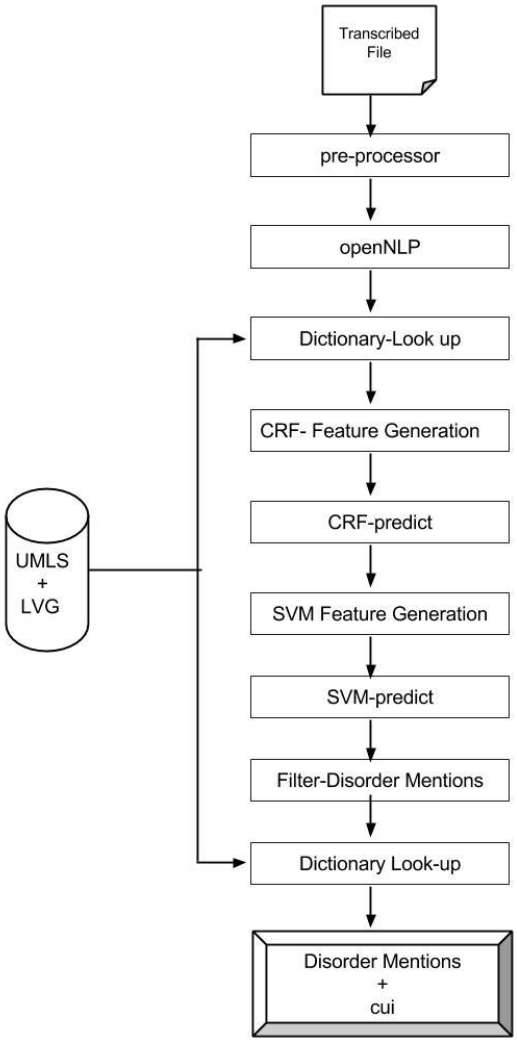

Figure 1: System Workflow

3.1.1 Pre-processing

All the clinical documents used in this task were de-identified. So information related to hospital name, patient demographics, physician names/signatures, dates, places, and certain lab-data were converted into predefined patterns. These patterns were hindering the flow of natural language. As a result of it, we were unable to get accurate results for PoS tagging and chunking of the sentences. So we replaced all of these de-identified patterns with some text that appear more as natural language. There were also some headers and footers associated with all the documents, which were actually irrelevant to this task. Therefore all headers and footers were also removed at the pre-processing level.

3.1.2 openNLP

We have used openNLP[3] to perform basic NLP tasks like sentence detection, tokenizing, PoS tagging, chunking, parsing and stemming.

3.1.3 Dictionary Lookup

UMLS 2012AA dictionary with Lexical Variant Generator (LVG)[4] was used to perform dictionary lookup task. Even though the task was only about finding disorder mentions, we also identified entities like procedures, finding, lab data, medicine, anatomical site and medical devices to be used as features in our CRF model. This was helpful in decreasing the number of false positive. UMLS TUI (Type Unique Identifier) used for different entity type is described in Table 3. A rule-based approach on the output of the OpenNLP syntactic parser was used to detect possible modifiers for disorder mentions.

Type	Tui list
Disorder	T046,T047,T048,T049,T050,T191, T037,T019,T184
Anatomical Sites	T017,T021,T023,T024,T025,T026, T029,T030
Procedures	T059,T060,T061
Medicines	T200,T120,T110
Lab Data	T196,T119
Modifiers	Customized Dictionary
Findings	T033,T034,T041,T084,T032,T201, T053,T054

Table 3: Entity Types and their related TUI list from UMLS

3.1.4 CRF Feature Generation

The feature sets were divided into three categories.
1) Clinical Features

i) Section Headers: A clinical note is often divided into relevant segments called Section Headers. These section headers provide very useful information at the discourse level. Same section header can have multiple variants. For example History of Present Illness can also be written as HPI, HPIS, Brief History etc. We have created a dictionary of more than 550 different section headers and classified them into more than 40 hierarchical categories. But using only section header dictionary for classification can fetch many false

[3] https://opennlp.apache.org/
[4] http://lexsrv2.nlm.nih.gov/

positives. Section header always appears in a predefined similar sequences. So to remove these false positives, we have used a Hidden Markov Model(HMM) (Parth Pathak, et al, 2013). For this task, we have used unigram section header id as a feature for all the tokens in CRF.

ii) Dictionary Lookup: A binary feature was used for all the different entity types detected from UMLS dictionary from last pipeline.

iii) Abbreviations: Abbreviations Disambiguation is one of the most challenging tasks in clinical NLP. The primary reason for the same is a lack of dictionary which contains most of the valid list of abbreviations. For this task, we have used LRABR as base dictionary to find out all the possible abbreviations and on top of that, a binary SVM classifier was used to check if the given abbreviation has medical sense or not.

2) Textual Feature:

Snowball stemmer[5] was used to find out stem value of all the word tokens. Prefix and suffix of length 2 to 5 were also used as features. Different orthographic features like whole word capital, first char capital, numeric values, dates, words containing hyphen or slash, medical units (mg/gram/ltr etc.) were used as features.

3) Syntactic Features:

Different linguistic features like PoS tags and chunks for each token were used. We have also used head of the noun phrase as one of the feature which can be very helpful in detecting the type of an entity.

3.1.5 CRF toolkit

All the annotated data was converted into BIO sequencing format. CRF++[6] toolkit was used to train and predict the model.

3.1.6 SVM

SVM was used to check whether a relationship exists between two entities or not. For this purpose all the tokens between these two entities, their part of speech tags and chunks were used as features. Rules based on output of a syntactic parser were also used as a binary feature. Some orthographic features like all letter capital, contains colon (:), contains semi colon (;), were also used as features. LibSVM[7] was used to train as well as predict the model.

3.1.7 Dictionary Look-up for CUI detection

For a better mapping of the entities detected by NLP inside the given input text, we found it to be a better approach to divide the UMLS entities into various phrases. This was done semi-automatically by splitting the strings based on function words such as prepositions, particles and non-nominal word classes such as verbs, adjectives and adverbs. While most of the disorder entities in UMLS can be contained into a single noun phrase (NP) there are also quite a few that contain multiple NPs related with prepositional phrases (PPs), verb phrases (VPs) and adjectival phrases (ADJPs).

This task gave us a modified version of the UMLS disorder entities along with their CUIs. The following table (Table 4) gives a snapshot of what this customized UMLS dictionary looked like.

CUI	Text	P1	P2	P3
C0013132	Dribbling from mouth	Dribbling	from	mouth
C0014591	Bleeding from nose	Bleeding	from	nose
C0029163	Hemorrhage from mouth	Hemorrhage	from	mouth
C0392685	Chest pain at rest	Chest pain	at	rest
C0269678	Fatigue during pregnancy	Fatigue	during	pregnancy

Table 4: An example of the modified UMLS disorder entities split as per their linguistic phrase types

Our dictionary look-up algorithm used this customized UMLS dictionary as resource to find the entities and assign the right CUIs.

4 Results & Error Analysis

4.1 Evaluation Calculations

The evaluation measures for Task A are Precision, Recall and F-Meas, defined as:

Precision = $\frac{TP}{FP+TP}$

Recall = $\frac{TP}{TP+FN}$

[5] http://snowball.tartarus.org/

[6] http://crfpp.googlecode.com/

[7] http://www.csie.ntu.edu.tw/\~cjlin/libsvm/

F-measure = $\frac{2*Precision*Recall}{Precision+Recall}$

where
TP = Disorder mention span matches with gold standard
FP = Disorder mention span detected by the system was not present in the gold standard;
FN = Disorder mention span was present in the gold standard but system was not able detect it

In Task B, the Accuracy was defined as the number of pre-annotated spans with correctly generated code divided by the total number of pre-annotated spans.

Strict Accuracy = $\frac{Total\ correct\ CUIs}{Total\ annotation\ in\ gold\ standard}$

Relaxed Accuracy = $\frac{Total\ correct\ CUIs}{Total\ span\ detected\ by\ system}$

4.2 System Accuracy

The system results were calculated on two different runs. For the first evaluation, only training data was used for the training purpose while for the second evaluation, both the training as well as the development data sets were used for training purpose. The results for Task A and B are as follows:

	Precision	Recall	F-Meas
Strict (T)	0.750	0.682	0.714
Relaxed (T)	0.915	0.827	0.869
Strict (T+D)	0.770	0.740	0.755
Relaxed (T+D)	0.911	0.887	0.899

Table 5: Task-A Results

where T= Training Data set
D= Development Data set

4.3 Error Analysis

Error Analysis on training data revealed that for Task-A our system got poor results in detecting non-contiguous disjoint entities. Our system also performed very poorly in identifying abbreviations and misspelled entities. We also observed that the accuracy of the part of speech tagger and the chunker also contributes a lot towards the final outcome. For Task-B, we got many false positives. Many CUIs which we identified from the UMLS were not actually annotated.

	Accuracy
Strict (T)	0.599
Relaxed (T)	0.878
Strict (T+D)	0.643
Relaxed (T+D)	0.868

Table 6: Task-B Results

5 Conclusion

In this paper we have proposed a CRF and SVM based hybrid approach to find Disorder mentions from a given clinical text and a novel dictionary look-up approach for discovering CUIs from UMLS meta-thesaurus. Our system did produce competitive results and was third best among the participants of this task. In future, we would like to explore semi-supervised learning approaches to take advantage of large amount of available unannotated free clinical text.

References

Savova, Guergana K., James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010. *Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications.* Journal of the American Medical Informatics Association 17, no. 5 (2010): 507-513.

Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. 1994. *A general natural-language text processor for clinical radiology.* J Am Med Inform Assoc 1994 Mar-Apr;1(2):16174. [PubMed:7719797]

Aronson, Alan R. 2001. *Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.* In Proceedings of the AMIA Symposium, p. 17. American Medical Informatics Association, 2001.

Wang, Yefeng, and Jon Patrick. 2009. *Cascading classifiers for named entity recognition in clinical notes.* In Proceedings of the workshop on biomedical information extraction, pp. 42-49. Association for Computational Linguistics, 2009.

Suominen, Hanna, Sanna Salanter, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan 2013 *Overview of the ShARe/CLEF eHealth evaluation lab 2013.* In Information Access Evaluation. Multilinguality, Multimodality, and Visualization, pp. 212-231. Springer Berlin Heidelberg, 2013.

. ””

Parth Pathak, Raxit Goswami, Gautam Joshi, Pinal Patel, and Amrish Patel. 2013 *CRF-based Clinical Named Entity Recognition using clinical Features*