ezDI: A Supervised NLP System for Clinical Narrative Analysis

Parth Pathak, Pinal Patel, Vishal Panchal, Sagar Soni,
Kinjal Dani, Narayan Choudhary, Amrish Patel
ezDI, LLC.
{parth.p, pinal.p, vishal.p, sagar.s,
kinjal.d, narayan.c, amrish.p} @ezdi.us

Abstract

This paper describes the approach used by ezDI at the SemEval 2015 Task-14: "Analysis of Clinical Text". The task was divided into two embedded tasks. Task-1 required determining disorder boundaries (including the discontiguous ones) from a given set of clinical notes and normalizing the disorders by assigning a unique CUI from the UMLS/SNOMEDCT[1]. Task-2 was about finding different type of modifiers for given disorder mention. Task-2 was divided further into two subtasks. In subtask-2a, gold set of disorder was already provided and system needed to just fill modifier types into the pre-specified slots. Subtask 2b did not provide any gold set of disorders and both the disorders and its related modifiers are to be identified by the system itself. In Task-1 our system was ranked first with F-score of 0.757 for strict evaluation and 0.788 for relaxed evaluation. In both Task-2a and 2b our system was placed second with weighted F-score of 0.88 and 0.795 respectively.

1 Introduction

Extracting medical information from clinical natural text has gained a lot of attraction over the past few years. Approximately 80% of patient related information resides under unstructured transcribed text. Amount of this unstructured text is increasing constantly and automated methods of extracting crucial information is of paramount interest to health care informatics industry. Task-14 of SemEval 2015 on "analysis of clinical text" addresses the same concern.

Task-14 of SemEval-2015 was in continuation of the 2013 ShaRe/CLEF Task-1 (Suominen, H. et al., 2013) and task-7 of SemEval 2014. The task was divided into two parts. In continuation of last year, task-1 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). This year one additional task (Task-2) of disorder modifier slot filing was added. Task-2 was further subdivided into two parts. In subtask-2a, a gold set of disorder mentions was provided and the participants had to only fill the pre-specified slots with the normalized modifiers. In task 2b, no gold set of disorder mentions was provided. Figure1 provides detailed overview about task 1 and 2.

Clinical NLP has evolved a lot in the tasks related to medical entity detection. NLP systems like cTAKES (Savova, Guergana K., et al., 2010), MetaMap (A. Aronson, 2001) and MedLEE (C. Friedman et al., 1994) have focused on rule based and dictionary look-up approaches for thid task. Recently a few attempts have been made to use supervised and semi-supervised learning models. In 2009, Yefang Wang (Wang et al., 2009) used cascading classifiers on manually annotated data and achieved around 83.2% accuracy. In 2010, i2b2 shared task challenge focused on finding test, treatment and problem mentions from clinical document. From 2013 on-words, entity detection task is regularly featuring in Share/CLEF and SemEval tasks.

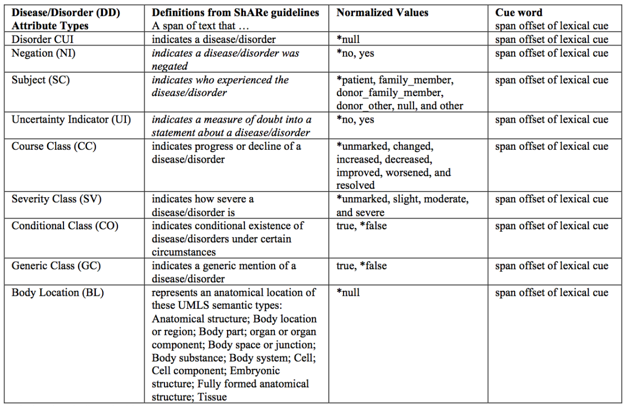
Tasks related to modifier slot filling are relatively

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

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 412–416,
Denver, Colorado, June 4-5, 2015.

new and no extensive research has been done yet. However for negation modifier, negEx (Chapman et al., 2011) or various other variants of negEx have been used in the last 10 years. These are keyword based dictionary look-up algorithms, but still gives around 92% of accuracy. However, these algorithms are not scalable because there is no proper mechanism defined to detect boundary for given negated keyword. In 2010 i2b2 challenge, there was a separate task for detecting 5 categories of negation. Systems used in this task showcase various statistical approaches and the accuracy numbers were in the range of 90 to 93%.

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



Disease/Disorder (DD) Attribute Types	**Definitions from ShARe guidelines** A span of text that …	**Normalized Values**	**Cue word** span offset of lexical cue
Disorder CUI	indicates a disease/disorder	*null	span offset of lexical cue
Negation (NI)	*indicates a disease/disorder was negated*	*no, yes	span offset of lexical cue
Subject (SC)	*indicates who experienced the disease/disorder*	*patient, family_member, donor_family_member, donor_other, null, and other	span offset of lexical cue
Uncertainty Indicator (UI)	*indicates a measure of doubt into a statement about a disease/disorder*	*no, yes	span offset of lexical cue
Course Class (CC)	indicates progress or decline of a disease/disorder	*unmarked, changed, increased, decreased, improved, worsened, and resolved	span offset of lexical cue
Severity Class (SV)	indicates how severe a disease/disorder is	*unmarked, slight, moderate, and severe	span offset of lexical cue
Conditional Class (CO)	indicates conditional existence of disease/disorders under certain circumstances	true, *false	span offset of lexical cue
Generic Class (GC)	indicates a generic mention of a disease/disorder	true, *false	span offset of lexical cue
Body Location (BL)	represents an anatomical location of these UMLS semantic types: Anatomical structure; Body location or region; Body part; organ or organ component; Body space or junction; Body substance; Body system; Cell; Cell component; Embryonic structure; Fully formed anatomical structure; Tissue	*null	span offset of lexical cue

Default values indicated with *

Figure 1: Task-2 with Examples.

2 Data Set

The SemEval-2015 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. Some other disorders which were not present in the UMLS were marked as CUI-less. The training and development data sets of the previous year's task were combined to be used as training set (298 documents) while the test data set of the previous year was used as development set. There were 100 documents used as test data set. Same set of 4 hundred thousand unlabelled documents were added to encourage use of unsupervised learning methods.

3 Disorder Detection and Normalization

For Task-1 our system was very similar to the system we developed last year (Pathak, et al, 2014). Entity detection task was converted into sequence labelling task using BIO format. A Conditional Random Fields (CRF) was used to detect continuous entity using CRF++[3] toolkit. To detect discontiguous entities, a binary SVM classifier was used to detect whether relationship existed between two disorder mentions or not. For contiguous entity detection task, our feature set was very similar to the one we used last year:

- Standard features like bag of words (for window +2 to -2), word stemmer (snowball stemmer) [4], prefix and suffix of length 1 to 5.
- Orthographic features like word contains digit, contains slash, contains special character and word shape (ezDI becomes aA).
- Grammatical features like parts of speech (PoS) tags for which we used an internally developed PoS tagger (Choudhary et al. , 2014), chunk (using Charniak's parser (Charniak and Johnson , 2005)) and head of noun and verb phrases.
- Dictionary look-up matches for window +2 to -2, stop words
- Section header and document type information and sentence cluster id

Support Vector Machine (LibSVM[5]) was used to identify disjoint entities. For all the possible combination of entities within a sentence, we ran a binary SVM classifier to find whether a relationship existed between those two entities or not. Feature set consisted of following features:

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

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

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

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

- Bag of words, PoS tags and chunk labels for all the tokens appearing in between two entities.

- Few simple rules were implemented on Charniak parse output to find relationship between two entities. A binary feature was used stating whether relationship was found using rules or not.

- Position of preposition, conjunction, main verb and special characters like colon (:), hyphen (-) and semi colon (;) in the context of the first entity.

- Binary feature stating whether any of the detected entity contained head of a noun phrase.

This hybrid approach was very helpful in detecting disjoint entities. We got around 70% accuracy in detecting disjoint entities using this approach.

3.1 CUI Detection

CUI detection task was divided into three separate steps:

1) Direct dictionary search: In the first step, for each word found in an entity we found all of its lexical variants using LVG [6]. After that, for all the possible permutations we tried searching the string in the UMLS. If the string matched any UMLS entry, we associated the corresponding CUI with that entity.

2) Dictionary search on modified entities: 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 exercise gave us a modified version of the UMLS disorder entities along with their CUIs. 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 1: An example of the modified UMLS disorder entities split as per their linguistic phrase types.

3) String similarity algorithm: If an entity was not found even after the first two steps, then we generated a list of possible text span from UMLS which can possibly match with the given entity. After that, Levenshtein edit distance algorithm was used to find best string match. If the best string match was greater than a certain threshold value, the corresponding CUI was associated with the entity otherwise the entity was marked as ”CUI-less”.

4 Modifier Detection:

For this task we tried to develop a generic approach so that it can be applied to any type of modifier. We divided the task of modifier detection into two parts: 1) Modifier keywords detection 2) Relating detected keywords with entity.

1) Modifier keywords detection: For each modifier type, an extensive dictionary was prepared having different possible keywords with its normalized values. A simple dictionary look-up algorithm was used to calculate a baseline accuracy. On training data set, accuracy ranged from 60% to 85% for different modifier types. This baseline algorithm achieved great recall but much less precision. To counter this, we used CRF algorithm with common features like bag of words, stem value and other orthographic features. CRF helped significantly in improving precision for modifier keyword detection.

2) Relating detected modifier with entities: We

[6] http://lexsrv2.nlm.nih.gov/

treated this task similar to the task of finding relationship between two entities. So a binary classifier was used to check if relationship existed between a modifier keyword and an entity or not. Feature set consisted of: Bag of Words between entity and modifier keyword, PoS tags, a binary flag stating whether the modifier keyword and the entity appeared in the same chunk or not, relative position of entity and modifier, special characters appearing in the sentence, section header (for subject modifier type).

5 System Accuracy

For Task-1, 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 = $\frac{\#of\ CUIs\ with\ Exact\ span\ match}{Total\ annotation\ in\ gold\ standard}$

Relaxed Accuracy = $\frac{\#of\ CUIswithpartialspanmatch}{Total\ annotation\ in\ gold\ standard}$

Both training and development data sets were used for training purpose. We used only 1 run with above mentioned system set up. We were ranked first for this task with results shown in Table 3.

	Precision	Recall	Accuracy
Strict	0.783	0.732	0.757
Relaxed	0.815	0.761	0.787

Table 2: Task-1 Results.

For Task 2, weighted and unweighted accuracies were calculated. The unweighted accuracy is the average of the per-disorder unweighted accuracy over all the disorders in the test set. Each gold-standard slot value is pre-assigned a weight based on its prevalence in the training set. The weighted accuracy is the average of the per-disorder weighted accuracy over all the disorders in the test set.

Ranks for task-2 were given based on weighted accuracy. ezDI was ranked second in both Task-2a and Task-2b. The results were as given below:

	F	A	F*A	WA	F*WA
Task-2A	1	0.934	0.934	0.880	0.880
Task-2B	0.915	0.935	0.856	0.868	0.795

Table 3: Task-2 Results.

6 Error Analysis

Abbreviations and disjoint entities still cause a lot of error in CUI normalization task. Dictionary related features are still not very helpful. Accuracy decreases significantly if medical domain is changed. Probably more sophisticated approach will be required to fully utilize UMLS dictionary. There is still a lot to explore in modifier detection. Statistical approaches are still not out-performing baseline dictionary based approaches. Modifier boundary detection is still a bigger challenge to be solved.

7 Conclusion

In this paper we have proposed a CRF and SVM based hybrid approach to find disorder mentions from a given clinical text, a novel dictionary look-up approach for discovering CUIs from UMLS metathesaurus and a generic statistical approach for modifier slot filling. Our system did produce competitive results and was best among all the participants for task 1. In future, we would like to explore semi-supervised learning approaches to take advantage of large amount of available un-annotated free clinical text.

References

Aronson, Alan R. 2001. *Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.*

Chapman W, Bridewell W, Hanbury P, Cooper G, Buchanan B. 2001. *A simple algorithm for identifying negated findings and diseases in discharge summaries.*

Charniak, Eugene and Mark Johnson. 2005. *Coarse-to-Fine n-best Parsing and MaxEnt Discriminative Reranking.*

Choudhary, Narayan, Parth Pathak, Pinal Patel, Vishal Panchal. 2014. *Annotating a Large Representative Corpus of Clinical Notes for Parts of Speech.*

Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. 1994. *A general natural-language text processor for clinical radiology.*

Pathak, Parth, Pinal Patel, Vishal Panchal, Narayan Choudhary, Amrish Patel, Gautam Joshi. 2014. *ezDI: A Hybrid CRF and SVM based Model for Detecting and Encoding Disorder Mentions in Clinical Notes.*

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.*

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.*

Wang, Yefeng and Jon Patrick. 2009. *Cascading classifiers for named entity recognition in clinical notes.*