TeamHCMUS: Analysis of Clinical Text

Nghia Huynh
Faculty of Information Technology
University of Science, Ho Chi Minh City, Vietnam
huynhnghiavn@gmail.com

Quoc Ho
Faculty of Information Technology
University of Science, Ho Chi Minh City, Vienam
hbquoc@fit.hcmus.edu.vn

Abstract

We developed a system to participate in shared tasks on the analyzing clinical text. Our system approaches are both machine learning-based and rule-based. We applied the machine learning-based approach for Task 1: disorder identification, and the rule-based approach for Task 2: template slot filling for the disorder. In Task 1, we developed a supervised conditional random fields model that was based on a rich set of features, and used for predicting disorder mentions. In Task 2, we based on the dependency tree to build a rule set. This rule set was extracted from the training data and applied to fill values of disorder attribute types on the test data. The evaluation on the test data showed that our system achieved the F-score of 0.656 (0.685 in case of relaxed score) for Task 1 and the F*WA of 0.576 for Task 2A and the F*WA of 0.671 for Task 2B.

1 Introduction

SemEval-2015 Task 14 is a continuation of previous tasks such as: CLEF eHealth Evaluation Labs 2013[1] (Hanna Suominen et al., 2013), CLEF eHealth Evaluation Labs 2014[2] (Liadh Kelly et al., 2014), and SemEval-2014 task 7[3] (Sameer Pradhan et al., 2014). The aim of the tasks is to improve the methods of natural language processing (NLP) of the clinical domain and to widely introduce the clinical text processing to the community of NLP research.

The clinical narrative is abundant in mentions of clinical conditions, anatomical sites, medications and procedures. It is completely different from the newswire domain where text is dominated by mentions of countries, locations and people. Many surface forms represent the same concept. Unlike the general domain, in biomedicine which are rich lexical and ontology resources that can be leveraged when applications are built.

The SemEval-2015 Task 14 is split into two tasks: 1) Task 1 is disorder identification, and its goal is to recognize the span of disorder mentions, the named entity recognition, and the normalization to a unique CUI in a SNOMED-CT terminology in a set of clinical notes. The SNOMED-CT is a resource provided by the organizers for the normalization of Task 1; and 2) Task 2 is disorder slot filling; it focuses on identifying the normalized value for nine modifiers in a disorder mentioned in a clinical note: the CUI of the disorder (much similar to Task 1), as well as the potential attributes (e.g. negation indicator, subject, uncertainty indicator, course, severity, conditional, generic indicator, and body location). Participants can submit to either or both of the tasks. We participated in both tasks.

In this paper, we describe a combined machine learning and rule-based approach for the two tasks.

2 Our Approach

2.1 Data Analysis

[1] https://sites.google.com/site/shareclefehealth/
[2] http://clefehealth2014.dcu.ie/
[3] http://alt.qcri.org/semeval2014/task7/

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 370–374,
Denver, Colorado, June 4-5, 2015.

The Organizing Committee provided a data set including one on-set of training data (train) and one on developing data (devel). The training data set contains 298 files, including radiology reports, discharge summaries, and ECG/ECHO reports. The developing data set contains 133 files being discharged summaries.

Processing the data shows that there are 3 forms to represent disorder mentions: 1) disorder with a continuous bundle of words (*Form 1*); 2) disorder with two separated chunks (*Form 2*); 3) disorder with three separated chunks (*Form 3*). Figure 1 illustrates the three forms. The statistics of the appearing rate of disorder representable forms on the training and the developing data sets are shown in Table 1.

Data		Form 1	Form 2	Form 3	Totals
Train	#disorder	10077	1028	62	11167
	Percentage	*91.8%*	*7.9%*	*0.3%*	
Devel	#disorder	7374	608	16	7998
	Percentage	*92.2%*	*7.6%*	*0.2%*	

Table 1. The statistics of the number and percentage of each disorder expressed in the sets of training and developing data.

Form 1: "The rhythm appears to be *atrial fibrillation*."

Form 2: "The *left atrium* is moderately *dilated*."

Form 3: ***"****Heart*: VI systolic murmur, *irregular* rate and *rhythm."*

Figure 1. Examples of disorder representable forms.

The analysis results help us develop a more effective disorder extraction approach in solving problems.

2.2 Disorder Identification

In disorder identification, the system is based on the machine-learning approach, the set of training data is converted into a BIO format, in which each word is assigned into one of three labels: B means the beginning of a disorder, I means the inside of a disorder, and O means the outside of a disorder. These labels can be used for a disorder only when it has consecutive words (*Form 1*) and cannot work when the disorder has nonconsecutive words (*Form 2* or *Form 3*) as mentioned in Section 2.1. Therefore, we developed different strategies for the disorder forms with consecutive and nonconsecutive words. For the disorder with consecutive words, we labeled words using the traditional BIO. For discontinuous disorder mentions, we created two addition sets of tags: 1) {B2, I2} which is used to assign to the words of disorder with two separate chunks (*Form 2*); 2) {B3, I3} is used to label the disorder with 3 separate chunks (*Form 3*). Figure 2 shows some examples of labeling disorders with consecutive and nonconsecutive words using our new tagging sets. In this approach, we assigned one of seven tags {B, I, O, B2, I2, B3, I3} to each word. Thus, the disorder identification problem was converted into a classification problem to assign one of the seven labels to each word.

Form 1: "The/O rhythm/O appears/O to/O be/O *atrial/B fibrillation/I* ./O"

Form 2: "The/O *left/B2 atrium/I2* is/O moderately/O *dilated/I2* ./O"

Form 3: *"Heart/B3* :/O VI/O systolic/O murmur/O ,/O *irregular/I3* rate/O and/O *rhythm/I3 ./O"*

Figure 2. Examples of labeling for the consecutive and nonconsecutive disorder words.

The algorithms machine learning and feature set offered by Stanford Named Entity Recognizer[4] was used. The Stanford CoreNLP[5] was used for splitting sentences and tokenizers from the training and test data. Also, some simple rules were used for labeling disorder words, i.e. {B, B2, B3) labeled to the begin-token of disorders, and {I, I2, I3} labeled to the inside-tokens of disorders as indicated in Figure 2. The Stanford-NER tool and the feature set offered by the Stanford NLP were used to build a supervised conditional random fields model on the training data. Then, this model was used to assign a label to each token in the test data. Some of our rules

[4] http://nlp.stanford.edu/software/CRF-NER.shtml
[5] http://nlp.stanford.edu/software/corenlp.shtml

were built to identify disorders. For sentences, we identified each disorder in turn based on the label sets consisting of {B, I}, {B2, I2}, and {B3, I3}.

In disorder normalization to a unique CUI in the UMLS/SNOMED-CT terminology, we extracted a list of annotated disorder from the training and developing date with disorder entities and CUI. This list was a primary search source for each of the recognized disorder entities. When a disorder was not found on the list, we used the MetaMap[6] (Willie, 2013) and UMLS[7] to continue the search. Then, when the disorder was not defined as CUI, it was defined as "CUI–less".

2.3 Disorder Slot Filling

Huu Nghia Huynh et al. (2014) developed a system to participate in Task 2 of the CLEF eHealth Evaluation Labs 2014. They used the rule-based and machine learning methods for the task of disease/disorder template filling. The result of the system achieved the accuracy of 0.827.

Our system was developed based on the rule-based approach. The rules are based on the representation of the dependency tree. One rule is established when there is a path from the node containing *disorder* to the node containing *Cue word* on the dependency tree. Each of these attributes has a rule set and a handling difference because of the data representation. For example, to fill values for the Uncertainty Indicator (UI) attribute, in the segment as illustrated in Figure 3, there are three disorders *"Congestive heart failure", "Coronary artery disease"* and *"Aortic valve disease"* whose all of the cue words are **"Indication"**. This segment are split into 3 sentences as shown in Figure 4, and *Sent 2* and *Sent 3* lost the Cue word information. Then when we based on the dependency tree, it is impossible to determine Normalized Values for an attribute.

Indication: *Congestive heart failure. Coronary artery disease. Aortic valve disease.*

Figure 3. Example of a text segment in discharge summary.

Sent 1: **Indication**: *Congestive heart failure*.
Sent 2: *Coronary artery disease*.
Sent 3: *Aortic valve disease*.

Figure 4. Example of separating the text segment result.

The attributes of Negation Indicator, Subject Class, Uncertainty Indicator, Course Class, Sever-ity Class, Conditional Class, and Generic Class are processed with the same method as follows: From the training and developing data, the system extracts lists, including a list of disorders and trigger and lists of the *Normalized Values* and the *Cue word* for each attribute. Every trigger list consists of two columns: the first column contains the *Normalized Values*, and the 2nd column contains the *Cue Word* of the respective disorder slot. The lists of disorder and trigger are the input parameters to define the sets of rules based on the dependency tree. Figure 5 is the illustration of the dependency tree in the sentence "Gastric lavage shown maroon/ black but no fresh blood" in which "blood" is a disorder, "no" is the *Cue word* of "blood" and the *Normalized Value* of "blood" to be determined is "yes".

A rule is set up to the Negation Indicator attribute type as follows: ({relation = "**neg**"} {governor = "***blood***"} {dependent = "**no**"}) → ("***blood***": yes). Each attribute has its own separated rule set.

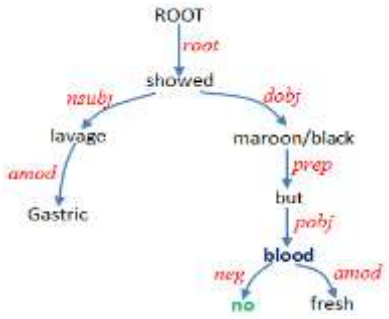

Figure 5. An example illustrates the dependency tree of the sentence "*Gastric lavage showed maroon/black but no fresh blood.*"

The disorder CUI attribute type is analyzed in the method similar to that of normalization of disorders mentioned above. For the Body Loca-

[6] http://metamap.nlm.nih.gov/JavaApi.shtml
[7] https://uts.nlm.nih.gov/home.html

tion attribute type, the system determines the *Cue word* candidates by searching the list of triggers and UMLS, and then uses the rule set to identify the *Cue word* related to the disorder.

3 Results

We used data that was provided by the organizers as training data for the system including 298 files (train) and 133 files (devel). The Organizing Committee provided the test data including 100 files (text) used to run Tasks 1 and 2b, followed by 100 files (pipe) used to run Task 2a. In task 2, there are two subtasks. In Task 2a, the gold-standard spans of disorder are given, and the participant has to fill the slots (including the CUI of the disorder). In Task 2b *End-to-end*: no gold-standard information is provided, and the participant has to (i) identify disorders (i.e. span recognition), and (ii) fill the slots for the disorders (including normalized disorders).

	Strict score	Relaxed score
Precision	0.680	0.711
Recall	0.633	0.662
F-score	0.656	0.685

Table 2. The system results of Task 1.

Accuracy	0.195
F*Accuracy	0.195
Wt_Accuracy	0.576
F*Wt_Accuracy	0.576

Table 3. The system results of Task 2a.

Accuracy	0.884
F*Accuracy	0.756
Wt_Accuracy	0.784
F*Wt_Accuracy	0.671

Table 4. The system results of Task 2b.

Attribute types	Weighted Accuracy
Body Location	0.603
Disorder CUI	0.801
Conditional Class	0.725
Course Class	0.851
Generic Class	0.904
Negation Indicator	0.935
Severity Class	0.843
Subject Class	0.931
Uncertainty Indicator	0.802

Table 5. The results of the attribute types in Task 2b.

Assessing the results in Task 2a, we made a mistake in filling out the default values for the slots of the disorders in the results submitted to the Organizing Committee. Therefore, the results are very low (see Table 3) and cannot reflect the effectiveness of our system.

The following metrics are computed with the F-measure for span identification: A true positive disorder span is defined as any overlap with a gold-standard span. If there are several predicted spans overlapping with a gold-standard one, then only one of them is chosen to be a true positive (the longest span), and the other predicted spans are considered as false positives.

#TP	5078
#FP	644
#FN	1070
Precision	88.7%
Recall	82.6%
F-score	85.6%

Table 6. The F-measure for span identification.

Table 6 illustrates the results obtained on the F-measure for span identification. On observing the results, a lot of predicted spans contain several tokens that were not part of disorders. If these tokens are removed, the results of span identification will be more accurate.

4 Discussion

The disorder identification task has a lot of challenges in the clinical domain. It was shown through the results in CLEF 2013 (Souminen,

H., et al., 2013), SemEval 2014 (Sameer Pradhan et al., 2014), and SemEval 2015.

Figure 6. Different representable forms of Disorders

In Section 2.1 we presented three representable forms of disorder in the clinical text. In addition, it shows the other representable forms as illustrated in Figure 6, and the different disorders sharing a word in the sentence. For example, the sentence 2 has 3 disorders containing the same word "***Allergies***".

The diversity and complexity of representation of disorders in clinical documents lead to a major challenge in the problem of extracting concepts in the clinical domain.

5 Conclusion

We described the system which realized the recognition, normalization and template filling of disorders in clinical docments. The system used the rule-based and machine learning-based approaches. The results of system will be able to serve a good foundation for our further research and propose enhancements to improve the efficiency for conceptual extraction problems. Specifically, we will study the proposal of more appropriate label sets for different representable forms of disorders as we presented in Sections 2.1 and 4, and conduct more pieces of research to supplement new features for disorder identification. In addition, we will propose a solution to remove several tokens which are not parts of disorder in the future.

References

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, Gareth J. F. Jones, Johannes Leveling9, Liadh Kelly, Lorraine Goeuriot, David Martinez, and Guido Zuccon. *Overview of the shARe/CLEF eHealth evaluation lab 2013*. In: Proceedings of ShARe/CLEF eHealth Evaluation Labs. (2013).

Huu Nghia Huynh and Bao Quoc Ho (2014). *A Rule-based Approach for Relation Extraction from Clinical Documents.* In Proceedings of Asian Conference 2014 on Information Systems, pp. 314-317.

Huu Nghia Huynh, Son Lam Vu, and Bao Quoc Ho (2014). *ShARe/CLEFeHealth: A Hybrid Approach for Task 2.* In Working Notes for CLEF 2014 Conference Sheffield, UK, pp. 103-110.

Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Tobias Schreck, Gondy Leroy, Danielle L. Mowery, Sumithra Velupillai, Wendy W. Chapman, David Martinez, Guido Zuccon, João Palotti (2014). *Overview of the ShARe/CLEF eHealth Evaluation Lab 2014.* In Information Access Evaluation. Multilinguality, Multimodality, and Interaction Lecture Notes in Computer Science Volume 8685, pp. 172-191.

Olivier Bodenreider and Alexa T. McCray (2003). *Exploring Semantic Groups through Visual Approaches.* Journal of Biomedical Informatics 36 (2003), pp. 414-432.

Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar and Guergana Savova (2014). *SemEval-2014 Task 7: Analysis of Clinical Text*. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp.54-62.

Willie Rogers (2013). *Installing and Running the Public Version of MetaMap*.