Medical Disorder Recognition with Structural Support Vector Machines

James Cogley, Nicola Stokes, and Joe Carthy

School of Computer Science and Informatics, University College Dublin, Dublin 4, Dublin, Ireland
james.cogley@ucdconnect.ie

Abstract. In this paper we present two systems that address the issues of disorder recognition and normalization submitted by the authors as defined by the CLEF/ShARe Evaluation Lab. The first approach to the tasks formed a baseline approach using the cTakes system. Our second approach leveraged Structural Support Vector Machines with an array of feature types including lexical, semantic and cluster based knowledge. The recognition differs from typical NER tasks in that disorder spans may be disjoint i.e. a disorder can be non-contiguous. To address this issue we introduce a new tag type to annotate tokens occurring between disorders.

Keywords: Support Vector Machines, Named Entity Recognition, Clinical Reports

1 Introduction

The CLEF/ShARe Evaluation Lab [1] provided a platform for comparative evaluation of clinical NLP about information retrieval technologies. The evaluation lab was comprised of three challenges:

1. Disorder recognition & normalization
2. Abbreviation recognition & normalization
3. Information Retrieval

For the purposes of this paper, we focus on the authors' submissions to the task of disorder recognition and normalization.

Our submissions had two goals: the first submission is based on the cTakes processing system, allowing us to gauge the suitability and performance of *ready-to-use* systems for the described tasks. The second submission is a machine-learning system built by the authors. This system allows for the identification of what feature sets aid the recognition task.

In the next section we present a brief overview of the task at the CLEF/ShARe Evaluation Lab.

2 Task

In this Section we present an overview of Task 1 at the ShARe/CLEF eHealth Evaluation Lab in which the authors participated. The dataset for this task comprised of 300 clinical reports (Discharge Summary, Radiology Report, Echo report) with disorder spans and normalized concepts from the UMLS Metathesaurus annotated. The training data comprised of 200 reports, with the released test data containing 100 reports.

Task 1 is divided into two subtasks: (a) disorder recognition, (b) disorder normalization. A Disorder is defined to an entity that may occur under the allowed groups in the UMLS Metathesaurus, as shown in Table 1. Though primarily a Named Entity Recognition (NER) task, it differs with previous clinical NER challenges, such as the i2b2 challenge. Firstly, an entity may be disjoint. That is to say that a recognized entity may or may not be a contiguous sequence of tokens, as is often a requirement in named entity challenges. For example, in Sentence 1., we see the concept text *Epstein's anomaly* separated by the token sequence *cardiac valve*.

1. Epstein's cardiac valve anomaly

The second unique aspect to the task requires the normalization of identified concept spans. For example, non-standard terminology may be used, or the concept is interrupted by a span of text, issues that may pose problems to information retrieval techniques. To perform normalization, spans are first identified. Following identification, spans are then mapped to ontology concepts, in the case of this challenge the UMLS metathesaurus is used. For the identified concept in sentence (1), it is mapped to the concept identifier C0013481 "Ebstein's anomaly". In the next section, we will present systems submitted by the authors that address these tasks.

Semantic Type
Congenital Abnormality
Acquired Abnormality
Injury or Poisoning
Pathologic Function
Disease or Syndrome
Mental or Behavioral Dysfunction
Cell or Molecular Dysfunction
Experimental Model of Disease
Anatomical Abnormality
Neoplastic Process
Signs and Symptoms

Table 1. Disorder: Semantic Groups

3 System Architecture

This section describes the two systems submitted to the CLEF/ShARe Task 1 in disorder recognition and normalization.

3.1 Preprocessing

Prior to recognizing disorders, the corpus must first be preprocessed. Firstly, documents were split into sentences using LingPipe[1] tools. Following sentence-splitting, the corpus is then passed through the cTakes system [2]. cTakes performs syntactic and semantic processing of the dataset as well as the recognition of named entities. The design of systems using this information is discussed in the following sections.

3.2 Baseline System : cTakes

cTakes facilitates information extraction from electronic medical health records. It is a comprehensive toolkit for processing clinical text, including abilities to detect named entities and map entities to CUI's in the UMLS metathesaurus. However, there are two issues that make cTakes an unsuitable candidate for disorder recognition and normalization. Firstly, cTakes cannot recognise disjoint entities. Secondly, as cTakes maps to all CUI's it raises issues in that we only require disorder normalization.

To achieve our aims, we apply simple post-processing rules on cTakes output in order to retrieve the required entities. Firstly, all recognised entities are filtered according to the allowed semantic groups that represent disorders in the UMLS Metathesaurus. In order to recognize disjoint disorders, our system performs a check that if another disorder with a matching CUI occurs within 10 tokens, those disorders are linked to create a disjoint entity.

3.3 Tagging with Structural SVM's

Despite the popularity of CRF's and other Markov approaches, Structural SVM's have been shown to achieve state-of-the-art performance with less training time on clinical datasets [3]. For this reason, the authors have designed an approach that leverages Structural SVM's as shown in Figure 1. Traditionally, tagging tasks such as Named Entity Recognition (NER) use the BIO (B-beginning, I-intermediate, O-outside) format. However, pre-submission experiments by the authors replicated results in the literature showing that improvements can be achieved using the BIESO (B-beginning, I-intermediate, E-end, S-single token concept, O-outside), particularly on long and short concepts. A key difference in the disorder recognition task and previous NER tasks is the allowance of disjoint concepts whereby disorders are not necessarily contiguous sequences of disorder tokens. To address this issue, the authors' use a modified BIESTO (T-beTween) tagging format that allows the tagging of tokens that occur between members of a disorder span as shown in Figure 2

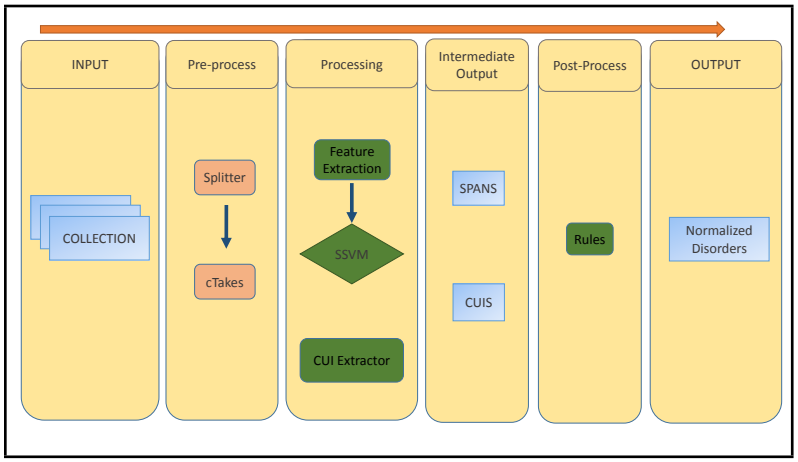

Fig. 1. Overview of Disorder Recognition and Normalization system using SSVM

The left atrium is moderately dilated.
O B I T T E

Fig. 2. Example of BIESTO tagging format

Features used in our machine-learning based system leveraged several types of information discovered in the text using the cTakes system and rules developed by the authors. Feature types included lexical, syntactic, semantic and clustering based features. To account for disjoint spans several features were introduced. Firstly, checks are performed to identify if matching CUI's occur in the same sentence. Features also analyze syntactic and part-of-speech information such as when the current token is a preposition to provide clues in identifying tokens that occur between spans. The normalization system is the same as described in Section 3.2 with the exception that the spans input to the system are those identified in the machine-learning system, rather than the cTakes system.

4 Evaluation

In this Section we present the evaluation metrics and results achieved by the systems described in the previous section.

4.1 Metrics

The disorder recognition systems are evaluated by the metrics precision (P), recall (R) and f-score (F):

$$P = \frac{TP}{TP + FP} \tag{1}$$

[1] `www.alias-i.com`

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

where TP is the number of true positives, FP is the number of false positives, FN is the number of false negatives.

For the task of recognising the spans of disorders in text, exact and inexact calculations of the above metrics are used. For an exact match, the begin and end offsets must match exactly. For an inexact calculation a candidate offset is counted as a True Positive if its spans overlap with the span of a gold standard annotation.

Accuracy is used as the evaluation metric for the normalization task. It is defined as follows:

$$Accuracy = \frac{CORRECT}{TOTAL} \tag{4}$$

where CORRECT is the number of disorders with correct span and CUI, TOTAL = Total number of disorders

Similar to the recognition task, there is both a strict and relaxed metric. For the strict metric, the gold-standard annotation is used as the count for Total number of disorders. In the relaxed metric, the system is evaluated with respect to only annotations detected by the system. In the next Section, we present the results of the systems submitted by the authors.

4.2 Results

This section provides the results of the authors submissions to the CLEF/ShARe Evaluation Lab. Our machine learning based approach to disorder recognition, `UCDCSI.1`, achieved competitive results across exact and inexact evaluations. The baseline cTakes system `UCDCSI.2` proved unsuccessful in recognizing disorder spans. Though not the worst performing system, it appears that a custom-built system for disjoint disorder recognition is required.

Examining the performance of the system on the exact metric showed some recurring errors. Firstly, given the strict nature of the metric, partial matches were penalised heavily. For example, the system recognised the instance *pulmonary hypertension* instead of the gold standard annotation *primary pulmonary hypertension*. However, issues such as this are addressed by the inexact metric.

System Name	P	R	F
UCDCSI.1	0.745	0.587	0.656
UCDCSI.2	0.268	0.175	0.212

Table 2. Disorder Recognition Exact-Spans

System Name	P	R	F
UCDCSI.1	0.922	0.758	0.832
UCDCSI.2	0.512	0.339	0.408

Table 3. Disorder Recognition Inexact-Spans

A second key cause of false positives would be the combination of a body-part and a modifier, such as *normalized gallbladder*.

This issue also extended to tests and treatments relating to a body part, such as *Liver function tests*. The final group of false positives were often negated disorders or those that featured some other assertion status, such as conditional or hypothetical. However, assertion status also generated issues with false negatives leading this to be a topic for further investigation. Typically, acronyms and abbreviations were a source of false negatives. While the system performed well in recognizing disjoint disorders, the performance weakened in detecting long disorders e.g. *mitral valve prolapse since adolescence who developed significant regurgitation* due to the long spaces between disjoint entities being atypical in the training set.

System Name	A
UCDCSI.1	0.299
UCDCSI.2	0.006

Table 4. Disorder Normalization Results

Both UCDCSI.1 and UCDCSI.2 relied on cTakes and post-processing rules to normalize concepts to CUI's in the UMLS Metathesaurus. However, this approach was far from effective, achieving moderate performance on the relaxed accuracy metric. This posits the idea that a free-standing module is required in order to correctly map identified concepts to CUI's. Analysis of the system's performance show that while cTakes may correctly map the CUI, its output produces several CUI's. Therefore, a more elegant set of post-processing rules may see the normalization process improve. For example, similarity measures may be used between an identified span and the textual representation of the identified CUI.

System Name	A
UCDCSI.1	0.509
UCDCSI.2	0.035

Table 5. Disorder Normalization Results (Relaxed)

5 Conclusions

In this paper, we have described a machine-learning based disorder recognition system using Structural SVM's and a novel BIESTO based tagging approach that facilitates the detection of disjoint entities. The system posted competitive results providing a solid foundation for future work. In particular, future work will focus on the normalization of recognised disorders. The approaches in this paper use frequency counts to map concept identifiers to disorders, future work may use semantic similarity measures to correctly identify normalized concepts.

Acknowledgements

This work was made possible through the Shared Annotated Resources (ShARe) project funded by the United States National Institutes of Health with grant number R01GM090187. We also wish to acknowledge the support of Science Foundation Ireland, who fund this research under grant number 10/RFP/CMS2836.

References

1. Suominen, H., Salantera, S., Velupillai, S.: Three shared tasks on clinical natural language processing. In: Proceedings of CLEF 2013. To appear. (2013)
2. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. Journal of the American Medical Informatics Association **17** (2010) 507–513
3. Tang, B., Cao, H., Wu, Y., Jiang, M., Xu, H.: Clinical entity recognition using structural support vector machines with rich features. In: Proceedings of the ACM sixth international workshop on Data and text mining in biomedical informatics. (2012)