LIST-LUX: Disorder Identification from Clinical Texts

Asma Ben Abacha, Aikaterini Karanasiou, Yassine Mrabet, and Julio Cesar Dos Reis*
Luxembourg Institute of Science and Technology (LIST)
29, avenue John F. Kennedy, L-1855 Kirchberg, Luxembourg
[asma.benabacha, aikaterini.karanasiou, yassine.mrabet]@list.lu
* Institute of Computing, University of Campinas
Av. Albert Einstein, 1251, Cidade Universitária Zeferino Vaz, 13083-852, Campinas, SP Brazil
julio.dosreis@ic.unicamp.br

Abstract

This paper describes our participation in task 14 of SemEval 2015. This task focuses on the analysis of clinical texts and includes: (i) the recognition of the span of a disorder mention and (ii) its normalization to a unique concept identifier in the UMLS/SNOMED-CT terminology. We propose a two-step approach which relies first on Conditional Random Fields to detect textual mentions of disorders using different lexical, syntactic, orthographic and semantic features such as ontologies and, second, on a similarity measure and SNOMED to determine the relevant CUI. We present and discuss the obtained results on the development corpus and the official test corpus.

1 Introduction

With the exponential growth of clinical texts, recognizing named entities becomes more and more important for several applications such as information retrieval, question answering or scientific analysis. The task of identifying mentions to medical concepts in free text and mapping these mentions to a knowledge base was recently proposed in ShARe/CLEF eHealth Evaluation Lab 2013 (Suominen et al., 2013).

The task 7 in SemEval 2014 (Pradhan et al., 2014) elaborates in that previous effort focusing on the recognition and normalization of named entity mentions belonging to the UMLS semantic group "Disorders". Similarly, task 14-1 of SemEval 2015[1] targets the identification of disorder mentions and their association to the relevant concept identifiers (CUI) in the UMLS/SNOMED_CT terminology. A disorder is normalized to "CUI-less" if the disorder mention is present, but there is no good equivalent CUI in UMLS/SNOMED-CT. Task 14-2b of SemEval 2015 specifically addresses *Disorder Slot Filling*. The aim is to identify the values of nine slots (negation indicator, subject, uncertainty indicator, course, severity, conditional, generic indicator and body location), given the span of disorder mentions from task 14-1.

In this paper we focus on task 1, *i.e.* disorder identification. In the following section we describe our approach to the detection of disorder mentions in clinical texts and their categorization with the relevant UMLS/SNOMED-CT CUI. In section 3 we present and discuss the obtained results on the development corpus and the official results before giving our concluding remarks in section 4.

2 Two-Step Approach for Disorder Identification

Our method includes two main steps: (1) the detection of disorder mentions using Conditional Random Fields (CRFs) and (2) the extraction of the associated CUI from SNOMED based on similarity measures. These two steps are described in more details in the following sections.

2.1 Step I - Disorder Mention Detection

The goal in this first step is to recognize the span of disorder mentions in a target clinical text. A mention can be a set of consecutive words, *e.g.* "***atrial***

[1] http://alt.qcri.org/semeval2015/task14/

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 427–432,
Denver, Colorado, June 4-5, 2015.

fibrillation", or disjoint, *e.g.* "***left atrium*** *is moderately* ***dilated***". In order to tackle the disjoint-mention problem, we annotated the data with the BIESTO format that is introduced by (Cogley et al., 2013).

2.1.1 BIESTO Labels

According to BIESTO format, the first word of a mention is tagged with B (beginning), the following words with I (inside), the last word with E (end) and the words between mention's words with T (beTween). The mentions that have one word are annotated as S (single) and the words that are not related to disorder mentions are annotated as O (outside). Furthermore, in the training and test corpus there are disorder mentions that end or start with the same word. In such case, when two serial B labels are followed by one E label, we consider two disorder mentions that start with different words and end with the same word. Similarly, if there is one B label followed by two different E labels, we consider two disorder terms that start with the same word and end with different words.

It is also observed that there is collision of BIESTO labels when one word exists into multiple disorder mentions and is annotated with different labels. In this case, we gather all the mentions which contain the common word and select the longest disorder mention (has the most words). If two mentions have the maximum length, the common word is annotated with two labels such as I/E.

Some examples of BIESTO labels are the following:

1. Disorder mentions that start with the same word, *e.g.*:
    - "The nasal septum deviates to the left with a rather large spur."
    - The nasal/B septum/I deviates/E to/T the/T left/T with/T a/T rather/T large/T spur/E.
    - "nasal septum deviates" and "nasal septum spur" are two disorder mentions with the same start word.
2. Collision between BIESTO labels, *e.g.*:
    - "osteophytes at C3/4 resulting in compression of the spinal cord with associated cord edema;"
    - Osteophytes/S at/O C3/O //O 4/O resulting/O in/O compression/B of/T the/T spinal/I/B cord/E/I with/T associated/T cord/T edema/E.
    - There are three disorder mentions: "Osteophytes" , "compression spinal cord" and "spinal cord edema".

2.1.2 CRF Algorithm

We use the Conditional Random Fields (CRFs) learning algorithm (Lafferty et al., 2001) in order to annotate the words with BIESTO labels. According to (McCallum and Li, 2003), suppose $x = \{x_1, x_2, x_3, ..., x_T\}$ is a set of input values (*e.g.* a sequence of words) and $s = \{s_1, s_2, s_3, ..., s_T\}$ is a set of states that are assigned to named entity labels, CRF estimates the conditional probability of a state sequence given an input sequence as follows:

$$P(s|x) = \frac{1}{Z} exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1}, s_t, x, t)\right)$$

where $1, ..., T$ represent the word positions, $1, ..., K$ represent the positions of the weighted features, the f_k represents the feature function and the λ_k is the weight of each feature function.

Using the CRF algorithm, the decision on a word's label can be influenced by the decision on the label of the preceding word. This dependency is taken into account in sequential models such as Hidden Markov Models (HMMs). However, the CRF model maximizes the conditional probability, unlike the HMM model which maximizes the joint probability. Therefore, the CRF model can use a number of features that are related to other words of the target texts in order to achieve better accuracy in its predictions. In our implementation we used the CRF++ tool[2].

2.1.3 Feature Set

In each experiment, we discard all the predicted disorder-mentions beyond 50 characters. In the last

[2]http://crfpp.googlecode.com/svn/trunk/doc/index.html

run, the "[**.....**]" and ":[** ..... **]" expressions as well as their lemmas and pos-tags were replaced by a sequence of "$".

We define a set of token and semantic features to train the CRF model.

Token features: The word, the part-of-speech tag (pos-tags) and the lemma; two tokens after and two tokens before the word, their lemmas and their pos-tags. We used StanfordTagger[3] to obtain the words of clinical texts as well as their lemmas and their part-of-speech tags.

StanfordTagger recognizes the word_1/word_2 token as one word. Since, many UMLS terms contain either the word_1 or the word_2, we separate the word_1/word_2 phrase into three words: word_1, / and word_2. For instance, given the following sentence: "*There is left lower lobe consolidation/volume loss.*", the system recognizes two disorder mentions that are: "*consolidation*" and "*volume loss*".

Linguistic and orthographic features: Indicating whether a word (i) is capitalized, (ii) contains digits, (iii) contains only lowercase characters without digits, the word length, suffixes and prefixes up to 4 characters.

2.2 Semantic Features

We use regular expressions to find the phrases which represent dates or time values (such as "2014-09-26", "4:07", "TUE", "Jan") and annotate them with the keyword DATE.

Stopwords (such as prepositions, conjunctions, articles) are annotated using a binary feature (yes/no). Precisely, if a word exists in the stopwords list[4], it is tagged with "YES", otherwise it is tagged with "NO".

Two features are derived from the Symptom Ontology[5] in order to annotate the words as SYMPTOM. We constructed a list of symptoms that contains the names of the ontology classes. If a word/phrase exists in the list of symptoms, then it is annotated as SYMPTOM. Since the names of ontology classes describe either a symptom or a group of symptoms, it is important to annotate only the names of symptoms. Consequently we added another feature which is the number of descendants for each class. The classes with no descendants (leaves) are likely to be symptoms and not a group of symptoms.

Following this same method, we annotate the words as DISEASES if they correspond to classes in the Human Disease Ontology[6].

One feature is derived from Human Development Anatomy Ontology[7] to annotate the words as anatomical_structure. We create a list of anatomical structures that contain the names of the ontology classes. If a word/phrase is in the list, it is tagged as Anatomical_Structure. We did not consider the number of descendants in this case because most of the names of ontology classes describe specific parts of the human body (anatomical structures).

Many phrases are frequent in clinical texts (*e.g.* headlines) and are not related to UMLS/SNOMED_CT terms. In order to improve the performance of the CRF algorithm, we gather and annotate them as OUTLINE. First, we extract all the phrases that end with colon and are located in the beginning of each sentence (such as "date of birth:", "review of symptoms:", "family history:") and we remove the phrases that contain digits (such as "Calcium 500 500 mg Tablet Sig:" and "[**2017-05-23**] 2:48 pm SWAB").

2.3 Step II - CUI Identification

In a second step we tackle the categorization of the detected disorder mentions with UMLS concept identifiers (CUI). The UMLS-Metathesaurus concept structure includes concept names, their identifiers, and some key characteristics of these concept names such as language and vocabulary source. In the *Rich Release Format* of the UMLS Metathesaurus, the important tables for this step are MRCONSO and MRSTY, which contain information about concepts and semantic types. The entire concept structure appears in MRCONSO while semantic types are obtained from the MRSTY.

A disorder mention is defined as any span of text that can be mapped to a concept in the SNOMED-

[3] http://nlp.stanford.edu/software/tagger.shtm

[4] http://www.ranks.nl/stopwords

[5] http://bioportal.bioontology.org/ontologies/SYMP

[6] http://www.obofoundry.org/cgi-bin/detail.cgi?id=disease_ontology

[7] http://www.obofoundry.org/cgi-bin/detail.cgi?id=human-dev-anat-abstract2

CT terminology, which belongs to the **Disorder semantic group**. A concept is in the Disorder semantic group if it belongs to one of 11 specific UMLS semantic types (87,412 concepts associated to disorders from 1,190,741 concepts of UMLS-2012AB) :

1. Congenital Abnormality (6130 concepts)
2. Acquired Abnormality (1746 concepts)
3. Injury or Poisoning (26607 concepts)
4. Pathologic Function (5115 concepts)
5. Disease or Syndrome (34213 concepts)
6. Mental or Behavioral Dysfunction (2710)
7. Cell or Molecular Dysfunction (383 concepts)
8. Experimental Model of Disease (3 concepts)
9. Anatomical Abnormality (1455 concepts)
10. Neoplastic Process (9050 concepts)
11. Sign or Symptom (2708 concepts)

We use SQL queries to construct our own table containing only disorders from the source "SNOMED" and related to the 11 semantic types (for a total of 348,760 rows). The proposed method then identifies the associated CUI for each disorder mention detected in step 1.

We start by performing an exact string comparison between the recognized disorder and the preferred terms and synonyms from the concepts of our table. If no exact match exists, we explore a similarity measure to calculate the relatedness between the detected mention and the available concepts. We use the *bigram* similarity measure following the observations of Cheatham and Hitzler (2013) on its suitability for ontology matching tasks. The selected CUI is the one with the highest similarity value. We fixed the word-based similarity threshold to 0.8 which led to the best results in our experiments (among different tested threshold values). If no exact match exists and all compared concepts have a similarity value under the threshold, the CUI-less class is associated to the detected mention.

3 Runs and Results

3.1 Evaluation Metrics

The results of our systems for task 14-1 are compared with the annotations of the gold-standard dataset using the F-measure, Precision and Recall metrics which are measured under strict and relaxed settings. In the strict setting, a disorder mention is correctly recognized, if its span and CUI code match exactly with a mention in the gold-standard dataset. In the relaxed setting, a disorder mention is correctly recognized if (i) there is an overlap with only one gold-standard mention from the same sentence, and (ii) the assigned CUI is correct.

In the following we present our results on the DEVelopment corpus (DEV) and the results on the official TEST corpus.

3.2 Experiments on the DEV Corpus

Table 1 presents the recall, precision and F-measure values for the strict and relaxed settings when different sets of features are used. More precisely, we consider the following sets:

- S1: Only Lexical features.
- S2: S1 + prefixes and suffixes.
- S3: S2 + labels of Symptoms ontology.
- S4: S3 + number of descendants for each symptom.
- S5: S4 + labels of Human Anatomy ontology.
- S6: S5 + number of descendants for each disease.

3.3 Configuration of the Submitted Runs

For the final evaluation we considered the two following sets of features: $Set_1 = \{$current word, 2 next words, 2 previous words lemmas, pos-tags, capital letters without digits, lower letters without digits, length of words, stop words, suffixes & prefixes [1,4], Dates/Time format $\}$ and $Set_2 = Set_1 \cup \{$labels from Symptom Ontology, number of descendants for each symptom, labels of Human Anatomy Ontology, labels from Human Disease Ontology, number of descendants for each disease $\}$ and we submitted 3 runs:

LIST-LUX, TASK1	strict_P	strict_R	**strict_F**	relax_P	relax_R	relax_F
S1	0.607	0.492	**0.543**	0.641	0.515	0.571
S2	0.601	0.544	**0.571**	0.633	0.568	0.599
S3	0.604	0.544	**0.572**	0.637	0.569	0.601
S4	0.604	0.544	**0.572**	0.638	0.570	0.602
S5	0.606	0.546	**0.575**	0.638	0.570	0.602
S6	0.609	0.547	**0.576**	0.641	0.572	0.604

Table 1: Results on the DEV corpus.

- Run 1: Feature Set_1, similarity threshold fixed to 0.8 for the CUI identification.
- Run 2: Feature Set_1, similarity threshold fixed to 0.83.
- Run 3: Feature Set_2, similarity threshold fixed to 0.8.

3.4 Official Results

Table 2 presents the final results on the TEST corpus[8]. When comparing the 3 runs we observe that increasing the similarity threshold had a slight negative impact on precision and a slight positive impact on recall. In a second observation, semantic features have a slight positive impact on both precision and recall which suggests their relevance, but also the need for larger ontologies and vocabularies.

Matching	Run	Precision	Recall	F-measure
Strict	1	0.649	0.577	0.611
	2	0.648	0.579	0.612
	3	0.649	0.580	0.613
Relaxed	1	0.677	0.602	0.637
	2	0.674	0.602	0.636
	3	0.675	0.603	0.637

Table 2: Task1: Official Results on the TEST corpus.

In order to evaluate the results in the second subtask, the metrics of F-measure, Precision, Recall, unweighted accuracy, weighted accuracy and per-slot weighted accuracy are estimated (*c.f.* table 3). Both unweighted and weighted accuracy are measures that show how well our system identifies all the slots for each disorder. The difference between them is that before estimating the weighted accuracy, each gold-standard slot value is assigned a weight based on its prevalence in the training corpus. The last metric is the Per-slot weighted accuracy that shows how well our system identifies the different values of each slot for all the disorders.

TASK2b	Run 1	Run 2	Run 3
F	0.884	0.882	0.881
A	0.865	0.866	0.866
F*A	0.765	0.763	0.763
WA	0.641	0.642	0.641
F*WA	**0.567**	**0.566**	**0.565**
BL	0.515	0.517	0.517
CUI	0.719	0.720	0.720
CND	0.496	0.500	0.497
COU	0.575	0.578	0.575
GEN	0.870	0.873	0.873
NEG	0.529	0.528	0.530
SEV	0.544	0.543	0.543
SUB	0.751	0.749	0.749
UNC	0.559	0.560	0.557

Table 3: Task 2b: Official Results on the TEST corpus.

3.5 Discussion

Table 4 presents the results of the first step (disorder detection) on the DEV corpus. It shows that F-measure decreased, in run 3, from 75,3% to 57,6% between mention detection (step 1) and CUI detection (step 2) in strict matching. Precision and Recall decreased with approximately the same factor. F-measure decreased, with a slightly higher factor in relaxed matching, from 86,1% to 60,4% between step 1 and step 2 (on the DEV corpus). Each matching setting shows a different estimation of the limitation related to similarity-based detection of CUI. This may be due to the additional noise when comparing partially-detected mentions with SNOMED

[8] http://alt.qcri.org/semeval2015/task14/index.php?id=results

labels and synonyms. Our similarity-based detection of CUI allowed reaching 57,6% F-measure on the DEV corpus and 61,3% F-measure on the TEST corpus (in strict matching, run 3), but it can still be enhanced further by taking into account additional features from the words surrounding the mentions and the concepts related to the candidate concepts in SNOMED (e.g. in the scope of global coherence maximization).

Matching	Run (Set)	P	R	F
Strict	R2 (S2)	0.792	0.717	0.752
	R3 (S6)	0.795	0.715	0.753
Relaxed	R2 (S2)	0.910	0.818	0.861
	R3 (S6)	0.913	0.814	0.861

Table 4: Task1: Results of the step 1 on the DEV corpus (disorder mention detection without CUI identification). P: Precision, R: Recall, F: F-measure.

4 Conclusion

In this article, we described our participation on two subtasks of the SemEval 2015 focused on disorder mention identification. We proposed a two-step approach suited to recognize spans of disorder mentions as a first step using a CRF learning algorithm with a set of features representing relevant aspects selected for the task. The method included a second step which accounted for the detection of adequate CUI from UMLS/SNOMEDCT concepts that might correspond to the recognized disorders from the target clinical texts. This research investigated the use of word-based similarity measures in the detection of CUI. The experiments running the method on two distinct corpora examined the influence of the defined features and configurations. Our approach based on CRF and similarity measures achieved 61.3% F-measure on the official TEST corpus. Using labels from ontology classes as semantic features was relevant for this task. In future work, we are planning to improve our CUI identification method. We are particularly considering the combination of supervised detection and categorization methods with semantic annotations obtained from unsupervised tools such as KODA(Mrabet et al., 2015) which allows annotating texts with both open-domain and domain-specific ontologies.

Acknowledgments

The last author was funded by São Paulo Research Foundation (FAPESP) (grant #2014/14890-0). We also want to acknowledge the efforts of the task organizers.

References

Michelle Cheatham and Pascal Hitzler. 2013. String similarity metrics for ontology alignment. In Harith Alani, Lalana Kagal, and *et. al.*, editors, *ISWC 2013*, volume 8219 of *LNCS*, pages 294–309.

James Cogley, Nicola Stokes, and Joe Carthy. 2013. Medical disorder recognition with structural support vector machines. In *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.*

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 188–191.

Yassine Mrabet, Claire Gardent, Muriel Foulonneau, Elena Simperl, and Eric Ras. 2015. Towards Knowledge Driven Annotation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, Austin, Texas, USA.

Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014. Semeval-2014 task 7: Analysis of clinical text. *SemEval 2014*, 199(99):54.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R South, Danielle L Mowery, Gareth JF Jones, et al. 2013. Overview of the share/clef ehealth evaluation lab 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 212–231.