IHS-RD-Belarus: Identification and Normalization of Disorder Concepts in Clinical Notes

Maryna Chernyshevich
IHS Inc. / IHS Global Belarus
131 Starovilenskaya St
220123, Minsk, Belarus
{Marina.Chernyshevich}@ihs.com

Vadim Stankevitch
IHS Inc. / IHS Global Belarus
131 Starovilenskaya St
220123, Minsk, Belarus
{Vadim.Stankevitch}@ihs.com

Abstract

This paper describes clinical disorder recognition and encoding system submitted by IHS R&D Belarus team at the SemEval-2015 shared task related to analysis of clinical texts. Our system is based on IHS Goldfire Linguistic Processor and uses a rich set of lexical, syntactic and semantic features. The proposed system consists of two components: a CRF-based approach to recognize disorder entities and empirical ranking to encode disorders to UMLS CUIs. Evaluation on the test data set showed that our system achieved the F-measure of 0.898 for entity recognition and the F-measure of 0.794 for UMLS CUI. The combined score for whole task is 0.690 (rank 17 out of 40 submissions).

1 Introduction

Named entity recognition (NER) is an information extraction task where the aim is to identify mentions of specific types of entities in text. This task has been one of the main focuses in the biomedical text mining research field, especially when applied to the scientific literature. Such efforts have led to the development of various tools for the recognition of diverse entities, including species names, genes and proteins, chemicals and drugs, anatomical concepts and diseases. These tools use methods based on dictionaries, rules, and machine learning, or a combination of those depending on the specificities and requirements of each concept type (Campos et al., 2013). After identifying entities occurring in texts, it is also relevant to disambiguate those entities and associate each occurrence with a specific concept, using a univocal identifier from a reference database such as Uniprot1 for proteins, or OMIM2 for genetic disorders. This is usually performed by matching the identified entities against a knowledge-base, possibly evaluating the textual context in which the entity occurred to identify the best matching concept.

In this paper, we describe a system (IHS_RD_Belarus in official results) developed to participate in the international shared task organized by the Conference on Semantic Evaluation Exercises (SemEval-2015) and focused on the analysis of clinical notes. This task is the repetition of task 7 at SemEval-2014 (Pradhan, et al., 2014) and aims at the recognition of entities belonging to the 'disorders' semantic group of the Unified Medical Language System (UMLS) (Bodenreider, 2004) and normalization of these entities to a specific UMLS Concept Unique Identifier (CUI). Specifically, the task definition required that concepts should only be normalized to CUIs that could be mapped to the SNOMED CT3 terminology.

2 System description

2.1 Dataset

The dataset for Tasks 1 consists of de-identified clinical notes of 4 different types (Discharge summary, ECG, Echo, Radiology) from MIMIC corpus (Lee et al., 2011). The organizer annotated 298 clinical notes with disorder entities on a predefined guideline and then mapped them to SNOMED-CT concepts represented by the UMLS CUIs. If a disorder entity cannot be found in SNOMED-CT, it was marked as "CUI-less". These notes were used as training dataset. The unlabelled notes are provided for exploring semi-supervised and unsupervised methods.

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 380–384,
Denver, Colorado, June 4-5, 2015.

Two types of disorder mentions are annotated: consecutive and discontiguous. The discontiguous dirorder mentions consist of multiple tokens with some distance between each other, for example, "*The left atrium is moderately dilated*".

Table 1 shows the counts of words, annotated disorders and unique CUIs in the training dataset.

	Train data
Documents	298
Words count	162,511
Disorder mentions	11,141
consecutive	10,050
discontiguous	1,091
Unique UMLS CUI	1,355
CUI-less entities	3,471

Table 1. Distribution of the training data.

2.2 Lexicon

The disorder lexicon was created using the UMLS Metathesaurus, where each disorder concept is represented by set of synonymous terms.

To satisfy the annotation guidelines, the concept identifiers (CUIs) were restricted to the 11 recommended disorder semantic types:

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

The disorder lexicon was enriched using automatically generated lists of synonymous words. For this purpose we used 3 techniques:

- lexical derivations, for example, *"optical, optically"*;
- synonymous words based on the Levenstein distance within a set of synonymous terms representing one UMLS disorder concept, for example, *"hyperchromasia, hyperchromatism, hyperchromia";*
- similar noun phrases suggested by our in-house autocorrection and autocompletion module that indexed UMLS terms, including correction of typing errors ("*carotic artery" = "carotid artery")* and similar terms (*"tick disease" = "tick-borne disease").*

2.3 Evaluation

Evaluation was to be carried out according to the following F-scores:

- *Strict F-score*: a predicted mention is considered a true positive if:
  1. its predicted span is exactly the same as for the gold-standard mention;
  2. the predicted CUI is correct.

  The predicted disorder is considered a false positive if the span is incorrect or the CUI is incorrect.
- *Relaxed F-score:* a predicted mention is a true positive if:
  1. there is any word overlap between the predicted mention span and the gold-standard span (both in the case of contiguous and discontinuous spans);
  2. the predicted CUI is correct.

  The predicted mention is a false positive if the span shares no words with the gold-standard span or the CUI is incorrect.

2.4 Disorder identification

We formulated disorder mention identification as a sequence labeling problem at token level and used Conditional Random Fields (CRF) (Lafferty, 2001). CRFs have shown empirical successes recently in named entity recognition (McCallum and Li, 2003), opinion target extraction (Chernyshevich, 2014), noun phrase segmentation (Sha and Pereira, 2003).

To facilitate feature generation for supervised CRF learning, sentences were pre-processed with IHS Goldfire Linguistic Processor that performs the following operations: word splitting, part-of-speech tagging, parsing, noun phrase extraction, semantic role labeling within expanded Subject-Action-Object (eSAO) relations (Todhunter et al., 2010). We removed all footers and headers, which are associated with the whole document and are irrelevant for the task. The notes are de-identified: the private data, e.g. names, data and places, are replaced by placeholders, for example, "`[**Location**]`". We replaced these placeholders with natural language expressions to assure correct POS-tagging and parsing.

Two separate CRF models were trained to identify consecutive and discontiguous disorder mentions with the same tagging scheme and same set of features.

2.4.1 CRF labels

We conducted several experiments with different tagging conventions and decided to use the ILO (Inside-Last-Outside) tagging scheme, where tag I represents the beginning and the inside token of an entity, L represents the last word of entity and O not a member of a disorder structure. The following is an example of our tagging for consecutive and discontiguous disorder mentions:

*The/**O** rhythm/**O** appears/**O** to/**O** be/**O** atrial/**I** fibrillation/**L***

*The/**O** left/**I** atrium/**I** is/**O** moderately/**O** dilated/**L***

The BIO (Begin-Inside-Outside) tagging scheme showed the classification accuracy lower by 5.5%.

2.4.2 Features

Given a sentence *s* and a token under consideration w_k, we define features over w_k and window of 5 tokens: w_{k-2}, w_{k-1}, w_k, w_{k+1}, w_{k+2}.

Token: This feature represents the string of the token w_k.

Context features: This feature has been used with a window of five tokens (the 2 tokens before and the 2 tokens after the target token). The surrounding words usually convey useful information about a token which help in predicting the correct tag for each token.

Part of speech: This feature represents the POS tag of the token w_k. It can provide some means of lexical disambiguation and help in determining the boundaries of instances.

Word letter case feature: This feature includes one of the three case tags for lowercase, uppercase and capitalized words correspondingly.

Letter n-grams: 3- and 4-letter n-grams starting and ending the token w_k.

Word frequency in out-of-domain corpus: we used social media texts as an out-of-domain corpus.

Part of a longer noun phrase: whether the word belongs to the same noun group as the next word.

Semantic category: This feature represents the semantic class to which the token w_k belongs, for example, body part, process, units of measure, drug, and animal being. We used two sources of semantic information: WordNet and the UMLS. The UMLS provides a set of semantic groups like anatomic terms, chemical substances and drugs, devices, disorders, etc. The WordNet was used to define semantic category of words not found in the UMLS. We selected the most representative nodes, for example, physical property, human, process etc. and all subordinate terms were assumed to belong to the appropriate category.

Document section: This unigram feature assigns the id of the section in which the token w_k belongs. Many clinical notes are divided into sections. These section headers provide very useful information, for example, the section "Past Medical History" or "Diagnosis" contains a lot of disorder mentions, while "Medications" do not. We created list of section headers, mapped to about 80 different unified names.

UMLS Features: We performed lookup in the disorder lexicon at two levels: word level and phrase level.

- The word-level feature represents the probability of a separate word to occur in a disorder mention. For this purpose, we collected all words contained in the UMLS disorders and calculated their probabilities of being a part of a disorder mention using the TF-IDF weighting. The TF of each word in the training set is calculated as the number of times the corresponding token appears in the UMLS disorder terminology. The IDF for each word is calculated from the number of unlabelled notes, which contain the word. These weighted metrics show how important the word is for disorder identification and help to exclude a lot of common words like frequent adjectives or conjunctions that often appear both in disorder terms and other terms.
- The phrase-level feature marks all phrases (with more than 2 words) that match a disorder term.

2.5 Disorder normalization

We propose a simple sieve-based algorithm that applies tiers of string matching for selecting the candidates with further candidates ranking.

2.5.1 Candidates selection

We applied following string matching rules to select candidate UMLS concepts for a disorder entity identified on the previous stage. Each rule assigns the score of confidence.

- **Exact match**: disorder and UMLS concept contain exactly the same extent text, excluding modifiers and determiners, with the same word order.

- **Relaxed match**: all informative words (excluding preposition, conjunctions, stop words etc.) from disorder are included in the UMLS concept.
- **Partial match**: at least one informative word from disorder is included in the UMLS concept.
- **Variants match**: all possible variants are generated for the disorder entity using synonyms, corrections and suggestions from our in-house autocorrection and autocompletion module and selected candidate UMLS concepts by relaxed matching rule.

2.5.2 Candidates ranking

All found candidate UMLS concepts were ranked on basis of a set of empirical parameters:

- score of match confidence;
- TF-IDF of the intersecting words;
- total number of disorder variants in the UMLS presenting the same CUI;
- number of times the UMLS concepts was already mentioned in this document;
- number of occurrences of the UMLS concept in the unlabelled corpus.

The top ranked UMLS concepts were selected as the system's output. If some concepts have the same ranking score, the first one by CUI number was selected.

2.6 Results and error analysis

The Table 2 summarizes the results separated by subtasks, disorder identification and disorder normalization, where the first column contains results obtained on development corpus and the second column shows the results on test corpus.

	Dev corpus	*Test corpus*
Disorder identification		
precision:	0.904	0.940
recall:	0.868	0.859
F1 measure:	0.886	0.898
Disorder normalization:		
accuracy:	0.794	0.794

Table 2: Separated results of disorder identification and normalization.

Our best performance on task 1 combining the disorder identification and normalization subtasks is shown in Table 3.

	Precision	*Recall*	*F1 measure*
Strict	0.722	0.662	0.690
Relaxed	0.746	0.684	0.714

Table 3: Combined result of disorder identification and normalization.

In this work we did not address the problem of discontiguous disorder mentions and correctly identified only about 10% of all discontiguous disorder mentions. Another source of errors are the one-, two-letters disorder acronyms, for example, "N", "V", "BM", etc. They remain untagged as diseases, as they may also refer to other entities, for example, chemicals.

As for disorder normalization task, the most challenging problem is the abbreviation disambiguation. The primary reason is a lack of abbreviations in UMLS terminology and their high ambiguity, for example, "AS" can refer to "Angelman Syndrome", "Aortic Stenosis", "Alzheimer Sclerosis" etc.

3 Conclusion

In this paper, we presented a clinical analysis system designed for participation in Task 1a of the SemEval 2015 Task 14 challenge. Our system performance was at 0.69 F-measure in the strict evaluation context and 0.714 F-measure in the relaxed evaluation context, obtaining a mid-range position. Our disorder recognition system presents good precision but performs worse in terms of recall, especially in discontiguous mentions identification. In order to improve our disorder normalization we plan to develop context similarity measures and improve the abbreviation disambiguation.

References

Andrew McCallum, Wei Li. Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. In the Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003.

Campos G., Vazquez A. I., Fernando R. L., K. Y. C., and S. Daniel, 2013. Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet. 7.

James Todhunter, Igor Sovpel and Dzianis Pastanohau. System and method for automatic semantic labeling of natural language texts. U.S. Patent 8 583 422, November 12, 2013.

John Lafferty, Andrew McCallum, and Fernando 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).

Joon Lee, Daniel J. Scott, Mauricio Villarroel, Gari D. Clifford, Mohammed Saeed and Roger G. Mark. Open-Access MIMIC-II Database for Intensive Care Research. In the Proceedings of the 33rd Annual International Conference of the IEEE EMBS, 2011.

F. Sha and F. Pereira. Shallow parsing with conditional random fields. In the proceedings of Human Language Technology/North American chapter of the Association for Computational Linguistics annual meeting (HLT/NAACL), 2003.

Maryna Chernyshevich. IHS R&D Belarus: Cross-domain extraction of product features using CRF. In the Proceedings of the International Workshop on Semantic Evaluation (SemEval), 2014.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. Nucleic Acids Research, 32:267–270.

Sameer Pradhan, Noemie Elhadad, Wendy Chapman, Suresh Manandhar and Guergana Savova. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation* (SemEval 2014), Dublin, Ireland.