Using Relations for Identification and Normalization of Disorders: Team CLEAR in the ShARe/CLEF 2013 eHealth Evaluation Lab

James Gung

University of Colorado, Department of Computer Science
Boulder, CO 80309
james.gung@colorado.edu

Abstract. This paper describes a system for span detection and normalization of disorder mentions in clinical notes as defined in Tasks 1a and 1b of the 2013 ShARe/CLEF eHealth Evaluation Lab [1]. We take a supervised learning, chunking-based approach to identifying disorder spans. In particular, our system introduces a method for identifying the spans of disjoint and overlapping disorder mentions using relation extraction and semantic role labeling (SRL). Our primary objective was to demonstrate the utility of relations for resolving the spans of disjoint disorder mentions. We used a CRF-based sequence labeler to extract initial disorder spans. Using these spans, we applied a locational relation extractor and SRL system to locate pairs of spans belonging to the same disorder mention. We used a dictionary-based approach to disorder normalization. Under strict evaluation for Task 1a, our system performed 3rd out of the 15 best performing systems for each team, achieving an F-measure of 0.687. For Task 1b, our system achieved a strict F-measure of 0.441. Our disjoint span resolution system significantly improved the performance of our system in both tasks, achieving a 5.5% increase over our baseline system in Task 1a and a 7.8% increase in Task 1b.

Keywords: clinical information extraction, named entity recognition, cTAKES, UMLS, semantic role labeling, relation extraction

1 Task Description

Task 1 of the 2013 ShARe/CLEF eHealth Evaluation Lab [1] had two components: span detection and normalization of disorders. The 2013 dataset consists of 300 discharge summaries, echo reports, ECG reports, and radiology reports taken from the MIMIC II database [2]. 100 of these were reserved for evaluation. Each document was annotated with disorders and corresponding CUIs (concept unique identifiers). Disorders were defined to be any span of text that could be mapped to a concept in SNOMED-CT terminology belonging to the Disorder semantic group [3]. Annotated disorder mentions often covered disjoint (non-contiguous) spans of text and in some cases overlapped. These characteristics pose many problems to traditional chunking-based approaches to NER and motivated much of the work in our approach.

2 Approach

We approach the task of identifying disorder spans as a supervised sequence labeling problem. Our system uses relation extraction and semantic role labeling to identify the spans of disjoint and overlapping disorder mentions. Our baseline system for the disorder span identification task uses a CRF (conditional random field) sequence labeler. The clinical Text Analysis and Knowledge Extraction System (cTAKES) is used to preprocess the data [4]. cTAKES is an open-source NLP system for information extraction from medical records built upon the UIMA framework. When extracting features, training and applying our sequence labeler, our system uses several components from ClearTK's machine learning module [5]. ClearTK is a framework for developing NLP applications also built upon Apache UIMA. CUI normalization is accomplished using a combination of rules and the cTAKES dictionary lookup algorithm.

2.1 Training Data

Before training the sequence labeler, we split up multi-span annotations into individual disorder annotations. For example, *right atrium* and *dilated* are annotated as disjoint spans within the disorder annotation *right atrium dilated* in the sentence “The right atrium is moderately dilated.” These disjoint spans are automatically annotated as independent disorder spans, while the original annotation is removed.

2.2 Processing Pipeline

The majority of preprocessing is accomplished using cTAKES components. Our baseline system's preprocessing pipeline consists of tokenization, sentence segmentation, part of speech tagging and NP-chunking. Finally, the cTAKES dictionary lookup module is applied over NPs found during chunking, providing entity mention annotations (drugs, diseases/disorders, signs/symptoms, anatomical sites, labs, procedures, and their associated CUIs).

2.3 Sequence Labeling

Our system uses the CRFsuite linear-chain CRF implementation for sequence labeling [6], wrapped into UIMA using ClearTK's machine learning module. After finding no significant improvements from the use of more advanced chunking schemes, we chose to apply the IOB (inside-outside-begin) chunking formalism.

Our system uses a combination of orthographic/lexical, syntactic and semantic features commonly used in named entity recognition. All orthographic/lexical and syntactic features are repeated in a window around the current token. A window of +/-2 tokens was found to provide the best performance on the training data. The normalized form of each token is extracted using a hand-crafted dictionary and is then used as a feature.

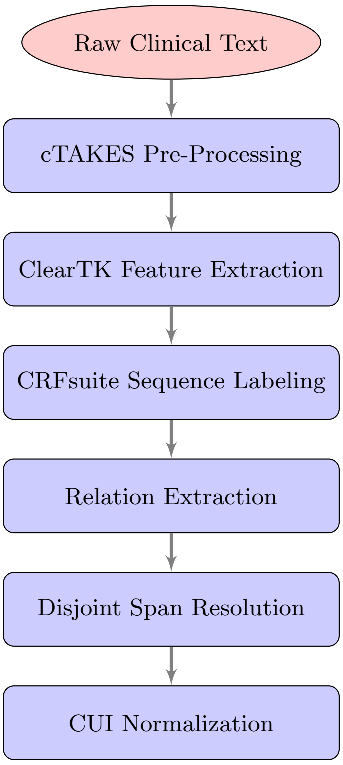

Fig. 1. Application Processing Pipeline for Disorder Span Detection and CUI Normalization

We also incorporate two domain-dependent features. *Discourse section* refers to the heading preceding the current section of text. With the observation that sections were generally colon-delimited in the task data, we naively identify the current discourse section as the text immediately preceding the previous colon. *Document type* is extracted from the filename of the document currently being processed. In the task data, the document type (ECHO_REPORT, ECG_REPORT, DISCHARGE_SUMMARY, or RADIOLOGY_ REPORT) could be found in the filename directly before the file extension (.txt).

Entity mentions extracted using the cTAKES the dictionary lookup module are used in the system as semantic features. Each entity type overlapping the span of the current token is included as a feature.

2.4 Disjoint Span Resolution

After identifying an initial set of disorder spans during sequence labeling, we apply our disjoint span resolution system to connect pairs of disjoint spans belonging to the same entities. This is accomplished using relations extracted with the

Table 1. Features used for Span Identification

Features	Example Text	Value(s)
CapitalType	CHF	ALL_CAPS
NumericType	15-20	SOME_DIGITS
Suffixes	pulmonary	ry, ary
Prefixes	pulmonary	pu, pull
CharacterCategoryPattern	Chronic	LuLl
NormalizedText	Abd	abdomen
PartOfSpeech	pain	NN
DisourseSection	dilated	LEFT ATRIUM
DocumentType	dilated	ECHO_REPORT
EntityType	atrium	ANATOMICAL_SITE

ClearNLP SRL module and the cTAKES relation extractor. Where locational relations extracted from cTAKES or predicate-argument relations extracted using SRL are found, we combine the disjoint spans corresponding to their constituents into single entities.

To identify locational relations, we use the cTAKES relation extractor module [4]. Using a binary LibSVM classifier trained on medical data, it identifies LocationOf relations (LocationOf[tumor, abdomen]) between pairs of entities identified with the cTAKES dictionary lookup algorithm. To identify semantic role relations, we use the ClearNLP semantic role labeler [7], a transition-based SRL system, also trained on medical data. The SRL model we applied does not explicitly identify adjectival predicates such as *dilated* in *the aortic root is moderately dilated.* Therefore, in order to capture disjoint spans with adjectival predicates, we looked for *be*-predicates that belonged in multiple SRL relations (such as [aortic root, is] and [dilated, is]) and treated the identified spans corresponding to their arguments as potential relations.

Finally, if a relation's constituent elements are contained within disorder spans from the initial set found during sequence labeling, we create a new, multi-span disorder mention from the corresponding spans. The old disorder spans are discarded.

2.5 CUI Normalization

To assign CUIs to the identified disorder spans, we apply a simple rule-based system in conjunction with the cTAKES dictionary lookup algorithm. We query cTAKES using the contents of each disorder mention as a lookup window. Because queries typically generate multiple results, we take several steps to filter candidate CUIs. First, we only consider identified annotations that cover the entire span. For example, *effusion* and *pericardial* in *pericardial effusion* are removed from consideration. Secondly, we only consider candidates that have TUIs (semantic type unique identifiers) belonging in the Disorder/Finding semantic groups. Finally, if no candidates are found, we mark the disorder span as CUI-less.

3 Results

Two runs were submitted for each task. In both tasks, CLEAR_NoRel corresponds to our baseline system without any disjoint span resolution while CLEAR_Rel incorporates our disjoint span resolution system. The tables below summarize the results for Task 1a and 1b using the SNOMED-CT 2011 gold standard. Also included in the tables are the results for the three best-performing systems out of all the teams excluding our system.

In Task 1a with strictly matching spans, CLEAR_Rel performed 3rd out of the 15 best performing systems for each team. Adding the disjoint span resolution system increased our baseline system's F-score by 0.036, a 5.5% improvement. For CUI normalization with strict evaluation, our system observed a 0.032 increase in F-score after adding disjoint span resolution, a 7.8% improvement.

Table 2. Task 1a Strict

System	Precision	Recall	F-score
CLEAR_NoRel	0.755	0.573	0.651
CLEAR_Rel	0.764	0.624	0.687
Mayo_1	0.800	0.573	0.668
NCBI_1	0.768	0.654	0.707
UTHealth_CCB_2	0.800	0.706	0.750

Table 3. Task 1a Relaxed

System	Precision	Recall	F-score
CLEAR_NoRel	0.937	0.705	0.804
CLEAR_Rel	0.929	0.759	0.836
Mayo_2	0.939	0.766	0.844
NCBI_2	0.904	0.805	0.852
UTHealth_CCB_2	0.925	0.827	0.873

Table 4. Task 1b Strict

System	Accuracy
CLEAR_NoRel	0.409
CLEAR_Rel	0.441
UTHealth_CCB_1	0.510
Mayo_2	0.546
NCBI_2	0.584

Table 5. Task 1b Relaxed

System	Accuracy
CLEAR_NoRel	0.714
CLEAR_Rel	0.706
Mayo_1	0.870
NCBI_1	0.890
AEHRC_1	0.939

4 Conclusions

We've introduced a system for disorder span detection and CUI normalization. An initial set of disorder spans are identified using a CRF-based IOB sequence labeler. CUI normalization is accomplished using the cTAKES dictionary lookup module and several simple rules. Using relations to resolve disjoint and overlapping spans significantly improves system performance in both disorder span detection and CUI normalization.

A more sophisticated system for normalizing abbreviations and acronyms would likely improve performance and make the system extensible to domains beyond ECGs, echo reports, radiographs and discharge summaries. Furthermore, CRFs have no built in capacity for capturing long distance dependencies [8]. This was visible in our error analysis where we found inconsistent treatment of identical spans of text in varying contexts. Intuitively, identical spans of text in the same domain should be consistently annotated the same way.

Although the cTAKES relation extractor and ClearNLP SRL system correctly identified many relations corresponding to entities with disjoint spans, many were also missed. For example, neither locational relations nor predicate-argument relations capture disjoint spans in coordination structures such as "right and left ventricles enlarged" in which *right ventricles enlarged* and *left ventricles enlarged* are both disjoint entities. Better results might be achieved by training a model to identify the relations that correspond to disjoint spans using the output from a simple dependency parse.

Acknowledgments. Many thanks to Martha Palmer, James Martin, Steven Bethard, Guergana Savova, and Sameer Pradhan for providing a wealth of ideas and advice in this task. This work was supported by award number 90TR002 provided by the ONC's Strategic Health IT Advanced Research Projects (SHARP). We gratefully acknowledge the efforts of all of the task organizers as well as the Shared Annotated Resources (ShARe) project funded by the United States National Institutes of Health with grant number R01GM090187.

References

1. H. Suominen, S. Salantera, S. Velupillai et al. Three Shared Tasks on Clinical Natural Language Processing. Proceedings of CLEF 2013. To appear.

2. G. Clifford, D. Scott, and M. Villarroel. User guide and documentation for the mimic ii database. 2012.
3. N. Elhadad, W. Chapman, T. O'Gorman, M. Palmer G. Savova. The ShARe Schema for the Syntactic and Semantic Annotation of Clinical Texts. Under Review.
4. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association : JAMIA*, 16(5):507–513, 2010.
5. P. V. Ogren, P. G. Wetzler, and S. J. Bethard. ClearTK: a framework for statistical natural language processing. In *Unstructured Information Management Architecture Workshop at the Conference of the German Society for Computational Linguistics and Language Technology*, 9 2009.
6. N. Okazaki. Crfsuite: a fast implementation of conditional random fields (crfs), 2007.
7. J. D. Choi and M. Palmer. Transition-based semantic role labeling using predicate argument clustering. *ACL HLT 2011*, page 37, 2011.
8. L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 147–155, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.