Disorder Concept Identification from Clinical Notes

An Experience with the ShARe/CLEF 2013 Challenge

Jung-wei Fan[1], Navdeep Sood[1], and Yang Huang[1*]

[1]Medical Informatics Group, Kaiser Permanente Southern California, 11995 El Camino Real, Suite 105, San Diego, CA 92130, USA

{jung-wei.x.fan, navdeep.x.sood, yang.x.huang}@kp.org

Abstract. We participated in both tasks 1a and 1b of the ShARe/CLEF 2013 NLP Challenge, where 1a was on detecting disorder concept boundaries and 1b was on assigning concept IDs to the entities from 1a. An existing NLP system developed at Kaiser Permanente was modified to output concepts that were close to the disorder definition of the Challenge. The core pipeline involved deterministic section detection, tokenization, sentence chunking, probabilistic POS tagging, rule-based phrase chunking, terminology look-up (using UMLS 2012AB), rule-based concept disambiguation and post-coordination. The system originally identifies findings (both normal and abnormal), procedures, anatomies, etc., and therefore a post-filter was created to subset the concepts with the source (SNOMED) and semantic types expected by the Challenge. A list of frequency-ranked CUIs was extracted from the training corpus to help break ties when multiple concepts were proposed on a single set of span. However, no retraining/customization was made to meet the boundary annotation preference specified in the challenge guidelines. Our best settings achieved an F-score of 0.503 (was 0.684 with relaxed boundary penalty) in task 1a, and best accuracy of 0.443 (was 0.865 on relaxed boundaries) in task 1b.

Keywords: medical language processing, concept boundary detection, concept normalization

1 Introduction

Natural language processing (NLP) has been an active and prolific subject in biomedical informatics [1, 2]. Organized open challenges sharing gold annotations constitute a critical driving force in biomedical NLP research and development, where annotated training corpora are scarce and valuable [3]. Aligned with the vision of facilitating clinical NLP, the ShARE/CLEF eHealth Evaluation Lab launched its first year (2013) challenge with tasks on extracting terms from clinical documents and normalizing them into standard terminology concepts [4]. For institutional interests, we participated specifically in task 1, which involved two sub-tasks. Task 1a was on detecting mention boundaries of concepts that belong to the Unified Medical Language System (UMLS) Disorders semantic group. A noteworthy feature of the challenge was that it

involved detecting concepts with discontinuous text spans. Task 1b was on normalizing each detected mention to a unique UMLS concept ID (CUI) that has SNOMED as one of its sources. We augmented an existing NLP system developed at Kaiser Permanente with special post-processors customized for the challenge. For task 1a, we achieved an F-score of 0.503 (and 0.684 with relaxed boundary penalty); for task 1b, our best accuracy was 0.443 (and 0.865 on relaxed boundaries).

2 Methods

For internal application interests, we developed an NLP system based on open-source tools (e.g., the Apache OpenNLP [5] and UIMA framework [6]). The system has core pipeline components that perform section identification, sentence chunking, tokenization, part-of-speech (POS) tagging, rule-based phrase chunking, concept look-up, sense disambiguation, and assertion classification. Due to the limitation and different focus of our current concept identification component, some modifications were made in order to better align with the challenge's requirements. The modifications are summarized as follows.

2.1 Identify concepts of discontinuous spans

Originally our concept identification could handle only concepts with a single continuous text span. To identify commonly observed discontinuous concepts in the target corpus, we manually analyzed our false negatives on the training set and composed concept post-coordinating rules. The rules apply a pairing template that searches within a sentence window for pre-specified concept A + concept B to infer a combined concept C. Table 1 shows some example rules.

Table 1. Examples of our concept post-coordinating rules

Component concepts	**Inferred concept**
C0225949 Leaflet of mitral valve C0205400 Thickened	C3164530 Thickened mitral leaflet
C0225844 Right atrial structure C0012359 Pathological Dilatation	C0344709 Right atrial dilatation
C0003501 Aortic valve structure C1285498 Vegetation	C0577870 Aortic valve vegetations
C0080310 Left Ventricular Function C0392756 Reduced	C1299337 Depression of left ventricular systolic function

2.2 Output SNOMED Disorders concepts

Our concept identification treats general findings and disorders as a single semantic class, and therefore requires modification to selectively output the disorder concepts defined by the challenge, which excludes non-symptomatic findings. Post-filter was created to select concepts that belong to the UMLS Disorders semantic group. Special logic was also created to check if an identified CUI has SNOMED-CT as one of its sources (our system included concepts of several other source vocabularies) and determine whether the concept ID should be the CUI or "CUI-less". If after the filtering there were still multiple concepts identified for a span (or a set of spans in discontinuous cases), we used concept prevalence computed from the training data to perform tie-breaking or just kept all the concepts when the tie-breaking failed (e.g., none of them ever occurred in the training data).

3 Results

Our best performance on task 1a is shown in Table 2. On task 1b we achieved a best accuracy of 0.443, and it was 0.865 when evaluated with boundary-relaxed (overlapping) concepts. The suboptimal performance was expected, since we did not customize our system settings to completely meet the challenge's preferences. For example, we considered T050 Experimental Model of Disease to be not useful and excluded its concepts, even though the semantic type belongs to the task-required Disorders semantic group. In addition, we did not agree with the boundary-marking approach used in the challenge's guidelines and therefore did not modify our system to behave likewise (see Discussion).

Table 2. Best performance our methods achieved on task 1a

	Recall	**Precision**	**F-score**
Strict boundary	0.512	0.494	0.503
Relaxed boundary	0.687	0.680	0.684

4 Discussion

We participated in the tasks 1a and 1b to get a feel about the quality of the gold annotation and assess its potential value for helping improve our NLP system. In general the human annotations offered insights on concepts we missed, especially the ones with discontinuous spans, which our system originally was not able to handle. However, there were a couple of fundamental properties in the annotation on which we held different perspective and therefore were not motivated to change our system to match. Below we discuss the properties in more detail.

4.1 Debatable boundary annotations in task 1a

It was not clear why the gold annotation tended to omit certain tokens in determining the concept spans, which oftentimes resulted in identifying less accurate concepts. For example, in 00211-027889-DISCHARGE_SUMMARY.txt the gold marked the substring "hematoma" as C0018944 Hematoma within "R groin small hematoma", which can actually be mapped to a more specific SNOMED-CT concept C0585249 Hematoma of groin. For such cases our system was double penalized for getting both a false negative and a false positive. Unexplainable token omissions were also observed in discontinuous spans: in 17582-104422-ECHO_REPORT.txt the gold selected three fragmented tokens "mitral", "leaflets", and "thickened" from the sentence "The mitral valve leaflets are mildly thickened" to represent a concept. However, if a system chose two alternatively viable spans "mitral valve leaflets" plus "thickened" to represent the same concept, it would be still penalized strict-boundary-wise. Since we did not see any obvious benefits in making our system reproduce such omissions, no customization was attempted accordingly.

4.2 Debatable concept ID annotations in task 1b

We suspected that the interplay among the constraints of only allowing Disorders concepts, only allowing SNOMED as source, and requiring a unique CUI assignment might have complicated the task unnecessarily. For example, Table 3 summarizes/comments on inconsistent CUI annotations observed in the training data for the expression "LV systolic function appears depressed". It can be argued whether such constraints are practical and have real benefit to downstream applications.

Table 3. Inconsistent CUI annotations for "LV systolic function appears depressed"

File name	Concept	Comment
13913-106200-ECHO_REPORT.txt	C1299337 Depression of left ventricular systolic function [T033 Finding]	This is supposed to be the best choice. However, it is not allowed according to the guidelines, which exclude T033.
03702-098383-ECHO_REPORT.txt	C1277187 Left ventricular systolic dysfunction [T046 Pathologic Function]	This is semantically close but not as accurate as the above. The annotator was tempted to use it most likely because T046 was guideline-allowed.
11801-104538-ECHO_REPORT.txt	CUI-less	This appears the majority and expected by the guidelines. However, it is compromising the fact that there is a perfect SNOMED concept available out there, i.e. the C1299337 above.

Besides, we believe the requirement of assigning a unique CUI to each concept can impose unjustifiable bias when there is actually more than one suitable choice. For

example, in 17522-024788-DISCHARGE_SUMMARY.txt the gold annotation mapped “chronic renal insufficiency” to C0022661 Kidney Failure, Chronic while our system chose C0403447 Chronic Kidney Insufficiency, which if not better, appears at least equally suitable.

Acknowledgement

This work is supported by the Shared Annotated Resources (ShARe) project funded by the United States National Institutes of Health: R01GM090187.

References

1. Spyns, P.: Natural language processing in medicine: an overview. Methods Inf Med. 35, 285–301 (1996).

2. Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F.: Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 128–44 (2008).

3. Chapman, W.W., Nadkarni, P.M., Hirschman, L., D’Avolio, L.W., Savova, G.K., Uzuner, O.: Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc. 18, 540–543 (2011).

4. Suominen, H., Velupillai, S.S.S., Chapman, W.W., Savova, G., Elhadad, N., Pradhan, S., South, B.R., Mowery, D., Jones, G.J.F., Leveling, J., Kelly, L., Goeuriot, L., Martinez, D., Zuccon, G.: Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. Proc of the CLEF 2013. To appear (2013).

5. Apache OpenNLP, http://opennlp.apache.org/.

6. Apache UIMA, http://uima.apache.org/.