Recognizing and Encoding Disorder Concepts in Clinical Text using Machine Learning and Vector Space Model

Buzhou Tang[1,2], Yonghui Wu[1], Min Jiang[1], Joshua C. Denny[3], and Hua Xu[1,*]

[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA
[2]Department of Computer Science, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong, China
[3]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA
{buzhou.tang, yonghui.wu, min.jiang, hua.xu}@uth.tmc.edu, josh.denny@vanderbilt.edu

Abstract. The ShARe/CLEF eHealth Evaluation Lab (SHEL) organized a challenge on natural language processing (NLP) and information retrieval (IR) in the medical domain in 2013. The first task of the 2013 ShARe/CLEF challenge was to extract disorder mention spans and their associated UMLS (Unified Medical Language System) concept unique identifiers (CUIs). We participated in Task 1 and developed a clinical disorder recognition and encoding system. The proposed system consists of two components: a machine learning-based approach to recognize disorder entities and a vector space model-based method to encode disorders to UMLS CUIs. The challenge organizers manually annotated disorder entities and corresponding UMLS CUIs in 298 clinical notes, of which 199 notes were used for training and 99 were for testing. Evaluation on the test data set showed that our system achieved the best F-measure of 0.750 for entity recognition (ranked first) and the highest F-measure of 0.514 for UMLS CUI encoding (ranked third), indicating the promise of the proposed approaches.

Keywords: medical language processing, natural language processing, named entity recognition, UMLS encoding, clinical concept extraction, conditional random fields, structured support vector machines, vector space model.

1 Introduction

Clinical natural language processing (NLP) has received great attention in recent years because it is critical to unlock information embedded in clinical documents in the secondary use of electronic health records (EHRs) data for clinical and translational research. Clinical concept extraction, which recognizes clinically relevant entities (e.g., diseases, drugs, labs etc.) in text and maps them to identifiers in standard vocabularies (e.g., Concept Unique Identifier (CUI) defined in Unified Medical Lan-

[*] corresponding author

guage System (UMLS) [1]), is one of the fundamental tasks in clinical NLP research. Many systems have been developed to extract clinical concepts from various types of clinical notes in last two decades. Earlier studies mainly focused on building symbolic NLP systems that are heavily based on domain knowledge (e.g., medical vocabularies). The representative systems include MedLEE [2], SymText/MPlus [3][4], MetaMap [5], KnowledgeMap [6], cTAKES[7], and HiTEX [8]. In the past few years, with the increasingly available annotated clinical corpora, researchers started to investigate the use of machine learning algorithms in clinical entity recognition. The Center for Informatics for Integrating Biology & the Beside (i2b2) has organized a few clinical NLP challenges to promote research in this field. In 2009, the i2b2 NLP challenge was to recognize medication-related concepts. Both rule-based and machine learning based methods as well as hybrid methods were developed by over twenty participating teams [9]. In the 2010 i2b2 clinical NLP challenge, organizers expanded clinical concepts from medication to problems, tests, and treatments. Most of systems were primarily based on machine learning algorithms in this challenge, likely due to the availability of large annotated datasets [10].

In 2013, the ShARe/CLEF eHealth Evaluation Lab (SHEL) organized three shared tasks on natural language processing (NLP) and information retrieval (IR): 1) clinical disorder extraction and encoding to Systematized Nomenclature Of Medicine Clinical Terms (SNOMED-CT), 2) acronym/abbreviation identification, and 3) retrieval of web pages based on queries generated when reading the clinical reports. The Task 1 on clinical disorder extraction is similar to the 2010 i2b2 challenge on clinical problem extraction. However, there are two major differences between these two tasks: 1) ShARe/CLEF task allowed disjoint entities, while 2010 i2b2 clinical problem extraction only dealt with entities of consecutive words; and 2) ShARe/CLEF task required mapping disorder entities to SNOMED-CT (using UMLS CUIs), which was not required in the 2010 i2b2 challenge.

In this paper, we describe our system for Task 1 of the 2013 ShARe/CLEF challenge. Our system consists of a machine learning based approach for disorder entity recognition and a Vector Space Model (VSM) based method for mapping extracted entities to SNOMED-CT codes. Evaluation by the organizers showed our system was top-ranked among all participating teams.

2 Methods

Fig. 1 shows the overview architecture of our systems for the first task of the ShARe/CLEF eHealth 2013 shared task. It is an end-to-end system of two components: disorder entity recognition and encoding. The first component consists of five modules. As the clinical narrative supplied by the organizer was not well formatted, we developed rule-based modules to detect the boundary of sentences and tokenize them for each note at first, and aligned the preprocessed note back to the original one at last. The other components were presented in the following sections in detailed.

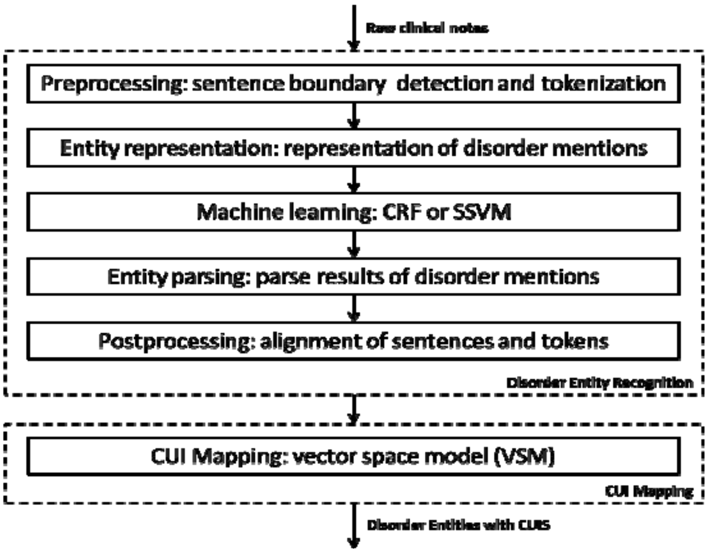

Raw clinical notes
Preprocessing: sentence boundary detection and tokenization
Entity representation: representation of disorder mentions
Machine learning: CRF or SSVM
Entity parsing: parse results of disorder mentions
Postprocessing: alignment of sentences and tokens
Disorder Entity Recognition
CUI Mapping: vector space model (VSM)
CUI Mapping
Disorder Entities with CUIS

Fig. 1. The overview architecture of our disorder concept extraction systems for the first task of the ShARe/CLEF eHealth 2013 shared task.

2.1 Dataset

The organizers collected 298 notes from different clinical encounters including radiology reports, discharge summaries, and ECG/ECHO reports. For each note, disorder entities were annotated based on a pre-defined guideline and then mapped to SNOMED-CT concepts represented by UMLS CUIs. If a disorder entity cannot be found in SNOMED-CT, it will be marked as “CUI-less”. The data set was divided into two parts: a training set of 199 notes that were used for system development, and a test set of 99 notes for evaluating systems. In the training set, 5811 disorder entities were annotated and mapped to 1007 unique CUIs or CUI-less. The test set contained 5340 disorder entities with 795 CUIs or CUI-less. Table 1 shows the counts of entities and CUIs in the training and test datasets.

2.2 Disorder entity recognition

In machine learning-based named entity recognition (NER) systems, annotated data are typically converted into a BIO format, where each word is assigned into one of three labels: B means beginning of an entity, I means inside an entity, and O means outside of an entity. Thus the NER problem is converted into a classification problem to assign one of the three labels to each word. As mentioned previously, one challenge of this task is that some disorder mentions (>10%) were disjoint, which could not be directly solved using the traditional BIO approach, which only works on entities with consecutive words. Therefore we developed different strategies for consecutive entities and disjoint entities. For consecutive disorder entities, we labeled words

Table 1. Statistics of the dataset.

Dataset	Type	#Note	#Mention	#CUI-less
Training	All	199	5816	1639
	ECHO	42	828	166
	RADIOLOGY	42	555	163
	DISCHARGE	61	3589	943
	ECG	54	193	90
Test	All	99	5340	1721
	ECHO	12	338	97
	RADIOLOGY	12	162	36
	DISCHARGE	75	4840	1588
	ECG	0	0	0

using traditional BIO tags. For disjoint entities, we created two additional sets of tags: 1) D{B, I} was used to label disjoint entity words that are not shared by multiple concepts (called non-head entity); and 2) H{B, I} was used to label head words that belonged to more than two disjoint concepts (called head entity). Figure 2 shows some examples of labeling consecutive and disjoint disorder entities using our new tagging sets. In this approach, we need to assign one of the seven labels {B, I, O, DB, DI, HB, HI} to each word. When converting labeled words to entities, we defined a few simple rules. For example, one rule for head words is "for each disjoint head entity, combine it with all other non-head entities to form final disorder entities".

Sentence 1: "The ***left atrium*** is ***dilated*** ."
Encoding: "The/O left/DB atrium/DI is/O dilated/DB ./O"
Sentence 2: "The ***aortic root*** and ***ascending aorta*** are moderately ***dilated*** ."
Encoding: "The/O aortic/DB root/DI and/O ascending/DB aorta/DI are/O moderately/O dilated/HB ./O"

Fig. 2. Examples of tagging for disjoint disorder entities.

We investigated two machine learning algorithms for disorder entity recognition. One is Conditional Random Fields (CRFs), which is a representative sequence labeling algorithm and is suitable for the NER problem. Another one is Structural Support Vector Machines (SSVMs), which was proposed by Tsochantaridis et al. [23] in 2005 for structural data, such as trees and sequences. It is an SVMs-based discriminative algorithm for structural prediction. Therefore, SSVMs combines the advantages of both CRFs and SVMs and is suitable for sequence labeling problems as well. CRFsuite (http://www.chokkan.org/software/crfsuite/) and SVM^{hmm} (http://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html) were used as implements of CRF and SSVM respectively.

For features, we used bag-of-word, part-of-speech (POS) from Stanford tagger (http://www-nlp.stanford.edu/software/tagger.shtml), type of notes, section information, word representation from Brown clustering [11] and random indexing [12], semantic categories of words based on UMLS [1] lookup, MetaMap [5], or cTAKEs [7] outputs. Most of features were the same as those used in our previous system for medical concept recognition [13][14][15][16].

2.3 Disorder entity encoding

We treated disorder entity encoding as a ranking problem, where each recognized disorder entity was considered as a query and candidates terms in UMLS as documents. The Vector Space Model (VSM) was used in this task. The process consists of two steps: 1) generate candidate CUIs from UMLS; and 2) rank candidate CUIs and then take the top ranked CUI as the system's output. We applied following criteria to select candidate CUIs from UMLS for a given disorder entity: the corresponding terms of a candidate CUI should contain all words in the disorder entity (except stop words). For each candidate CUI, a vector containing its words, weighted by term frequency–inverse document frequency (tf-idf) derived from entire UMLS/SNOMED-CT terms, was created. The cosine similarity between a disorder entity vector and a candidate CUI vector was calculated and used to rank candidate CUIs. The top ranked CUI was then selected as the correct CUI of the entity. In order to leverage the training data, we further built a limited VSM-based encoding system by using CUIs/terms and entities occurred in the training set only, instead of the entire UMLS. When processing the test set, we first determined whether an entity occurred in the training set or not. If it did, we used the limited VSM-based encoding system to predict the corresponding CUI. Otherwise, we used the general VSM-based encoding system that was built on entire UMLS.

2.4 Experiments and Evaluation

Our system was developed and trained using the training set (199 notes) and was evaluated using the test set (99 notes). All parameters of CRF and SSVM were optimized by 10-fold cross-validation on the training dataset. The performance of disorder entity recognition were evaluated by precision, recall and F-measure in both "strict" and "relaxed" modes, where "strict" refers that a concept is correctly recognized if and only if the starting and ending offsets of it is exactly same as a disorder mention in the gold standard, and "relaxed" refers that a disorder mention is correctly recognized as long as it overlaps with any disorder mention in the gold standard. For encoding of SNOMED-CT, all participating systems were evaluated using accuracy only, in "strict" and "relaxed" modes, as defined in [17][18].

3 Results

Table 2 shows the best performance of our system in the ShARe/CLEF eHealth 2013 shared task 1 as reported by the organizers, where "Pre", "Rec", "F" and "Acc" denote precision, recall, F-measure and accuracy respectively. For disorder entity recognition, the SSVM-based system outperformed CRF-based system, achieving the best

F-measures of 0.750 under “strict” criterion and 0.873 under “relaxed” criterion, ranked first in the challenge. For SNOMED encoding, our system achieved the best accuracy of 0.514, ranked third in the challenge.

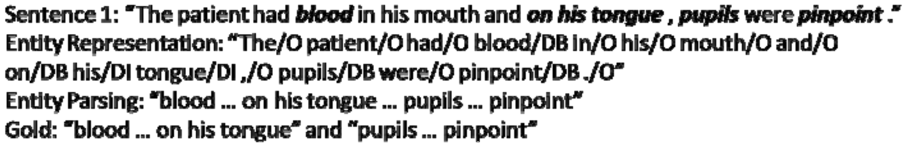
Table 2. The performance of our system for the ShARe/CLEF eHealth 2013 shared task 1.

Task	Strict				Relaxed			
	Pre	Rec	F	Acc	Pre	Rec	F	Acc
Task 1a (entity recognition)	0.800	**0.706**	**0.750**	-	0.925	**0.827**	**0.873**	-
Task 1b (SNOMED encoding)	-	-	-	**0.514**	-	-	-	**0.729**

4 Discussion

Although a number of existing clinical NLP systems such as MedLEE [2], MetaMap [5], KnowledgeMap [6], and cTAKES [7] can extract clinical concepts and map them to UMLS CUIs, it is difficult to compare the performance of these systems because there is a lack of publically available corpora with annotations of UMLS CUIs. The 2013 ShARe/CLEF eHealth shared task 1 provides such a benchmark dataset for clinical concept recognition and encoding, which is a significant contribution to the clinical NLP research. Furthermore, the best system in the challenge achieved an accuracy of 0.589 on encoding SNOME concepts, indicating it is still very challenging to develop general clinical NLP systems that can accurately recognize and encode clinical disorders to standard terminologies.

In this study, we developed a clinical disorder recognition and encoding system that combines a machine learning based approach for entity recognition and a VSM-based approach for UMLS concept mapping. Our system was top-ranked among all participating teams, indicating the promise of proposed approaches. However, there is still much room for further improvement. First, our proposed method for disjoint entity recognition has limitations. For example, if a sentence has multiple disjoint entities, our current simple rule-based strategies would not be able to resolve the ambiguity and will produce wrong combinations of disorder entities as shown in Fig 3, where there are two disorder entities in the given sentence: “blood … on his tongue” and “pupils … pinpoint”, which are represented by “blood/DB … on/DB his/DI tongue/DI” and “pupils/DB … pinpoint/DB” respectively, but parsed into one disorder entity “blood … on his tongue … pupils … pinpoint” by our strategies. Thus, more sophisticated methods for disjoint concept recognition should be investigated in future. In addition, our VSM-based method to map entities to UMLS CUIs is not optimal. When compared with the top ranked team on UMLS CUI mapping, our system achieved better performance on entity recognition, but lower accuracy on CUI mapping, indicating the weakness of our encoding step. A few possible aspects for further improvement are: 1) use other types of information as features for building vectors, such as context, type of notes, section information and so on; 2) explore other ranking algorithms such as Support Vector Machines [19], and 3) implement word sense disambiguation algorithms for ambiguous entities.



Sentence 1: "The patient had ***blood*** in his mouth and ***on his tongue , pupils*** were ***pinpoint*** ."
Entity Representation: "The/O patient/O had/O blood/DB in/O his/O mouth/O and/O on/DB his/DI tongue/DI ,/O pupils/DB were/O pinpoint/DB ./O"
Entity Parsing: "blood ... on his tongue ... pupils ... pinpoint"
Gold: "blood ... on his tongue" and "pupils ... pinpoint"

Fig. 3. Examples of entity parsing errors.

5 Conclusions

We developed a clinical disorder recognition and encoding system that consists of a machine learning-based approach to recognize disorder entities and a vector space model-based method to encode disorders to UMLS CUIs. Our entry based on this system was top-ranked in the 2013 ShARe/CLEF eHealth shared task 1, indicating the promise of our approaches. However, more investigations are needed in order to achieve satisfactory performance on extracting and encoding medical concepts in clinical text.

Acknowledge
This study is supported in part by grants from NLM R01LM010681, the Office of the National Coordinator for Health Information Technology 10510592, NCI 1R01CA141307, and NIGMS 1R01GM102282. We also thank the ShARe/CLEF eHealth shared task 2013 organizers, who were funded by the United States National Institutes of Health with grant (R01GM090187).

References

[1] "Unified Medical Language System (UMLS) - Home." [Online]. Available: http://www.nlm.nih.gov/research/umls/. [Accessed: 22-May-2013].

[2] C. Friedman, P. O. Alderson, J. H. Austin, J. J. Cimino, and S. B. Johnson, "A general natural-language text processor for clinical radiology.," *J Am Med Inform Assoc*, vol. 1, no. 2, pp. 161–174, 1994.

[3] S. B. Koehler, "SymText : a natural language understanding system for encoding free text medical data;," University of Utah;, 1998.

[4] L. M. Christensen, P. J. Haug, and M. Fiszman, "MPLUS: a probabilistic medical language understanding system," in *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain - Volume 3*, Stroudsburg, PA, USA, 2002, pp. 29–36.

[5] A. R. Aronson and F.-M. Lang, "An overview of MetaMap: historical perspective and recent advances," *J Am Med Inform Assoc*, vol. 17, no. 3, pp. 229–236, May 2010.

[6] J. C. Denny, P. R. Irani, F. H. Wehbe, J. D. Smithers, and A. Spickard, “The KnowledgeMap Project: Development of a Concept-Based Medical School Curriculum Database,” *AMIA Annu Symp Proc*, vol. 2003, pp. 195–199, 2003.

[7] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, “Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications,” *J Am Med Inform Assoc*, vol. 17, no. 5, pp. 507–513, Sep. 2010.

[8] Q. T. Zeng, S. Goryachev, S. Weiss, M. Sordo, S. N. Murphy, and R. Lazarus, “Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system,” *BMC Med Inform Decis Mak*, vol. 6, p. 30, 2006.

[9] O. Uzuner, I. Solti, and E. Cadag, “Extracting medication information from clinical text,” *J Am Med Inform Assoc*, vol. 17, no. 5, pp. 514–518, Oct. 2010.

[10] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, “2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text,” *J Am Med Inform Assoc*, vol. 18, no. 5, pp. 552–556, Oct. 2011.

[11] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, “Class-Based n-gram Models of Natural Language,” *Computational Linguistics*, vol. 18, pp. 467–479, 1992.

[12] K. Lund and C. Burgess, “Producing high-dimensional semantic spaces from lexical co-occurrence,” *Behavior Research Methods, Instruments, & Computers*, vol. 28, no. 2, pp. 203–208, Jun. 1996.

[13] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu, “A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries,” *J Am Med Inform Assoc*, vol. 18, no. 5, pp. 601–606, Oct. 2011.

[14] B. Tang, H. Cao, Y. Wu, M. Jiang, and H. Xu, “Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features,” *BMC Med Inform Decis Mak*, vol. 13 Suppl 1, p. S1, 2013.

[15] B. Tang, H. Cao, Y. Wu, M. Jiang, and H. Xu, “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*, New York, NY, USA, 2012, pp. 13–20.

[16] B. Tang, Y. Wu, M. Jiang, Y. Chen, J. C. Denny, and H. Xu, “A hybrid system for temporal information extraction from clinical text,” *J Am Med Inform Assoc*, Apr. 2013.

[17] H. Suominen, S. Salantera, S. Sanna, and et al, “Overview of the ShARe/CLEF eHealth Evaluation Lab 2013,” presented at the Proceedings of CLEF 2013, 2013, p. To appear.

[18] W. Chapman, G. Savova, and N. Elhadad, “ShARe/CLEF Shared Task 1 for boundary detection and normalization of SNOMED disorders,” presented at the Proceedings of CLEF 2013, 2013, p. To appear.

[19] T. Joachims, “Optimizing search engines using clickthrough data,” in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2002, pp. 133–142.