UTH_CCB: A Report for SemEval 2014 – Task 7 Analysis of Clinical Text

Yaoyun Zhang[1] Jingqi Wang[1] Buzhou Tang[2] Yonghui Wu[1] Min Jiang[1] Yukun Chen[3] Hua Xu[1*]

[1]University of Texas School of Biomedical Informatics at Houston Houston, TX, 77030, USA

[2]Harbin Institute of Technology Shenzhen Graduate School Shenzhen, 518055, China

[3]Vanderbilt University Department of Biomedical Informatics Nashville, TN, 37240, USA

{Yaoyun.Zhang, Yonghui.Wu, Min.Jiang, Hua.Xu} @uth.tmc.edu

tangbuzhou@gmail.com yukun.chen@Vanderbilt.Edu

Abstract

This work describes the participation of the University of Texas Health Science Center at Houston (UTHealth) team on the SemEval 2014 – Task 7 analysis of clinical text challenge. The task consisted of two subtasks: (1) disorder entity recognition, recognizing mentions of disorder concepts; (2) disorder entity encoding, mapping each mention to a unique Concept Unique Identifier (CUI) defined in Unified Medical Language System (UMLS). We developed three ensemble learning approaches for recognizing disorder entities and a Vector Space Model based method for encoding. Our approaches achieved top rank in both subtasks, with the best F measure of 0.813 for entity recognition and the best accuracy of 74.1% for encoding, indicating the proposed approaches are promising.

1 Introduction

In recent years, clinical natural language processing (NLP) has received great attention for its critical role in unlocking information embedded in clinical documents. Leveraging such information can facilitate the secondary use of electronic health record (EHR) data to promote clinical and translational research. Clinical entity recognition, which recognizes mentions of clinically relevant concepts (e.g., disorders, procedures, drugs etc.) in narratives, and clinical entity encoding, which maps the recognized entities to concepts in standard vocabularies (e.g., UMLS CUI (Bodenreider, 2004)), are among 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, ranging from early symbolic NLP systems heavily dependent on domain knowledge to machine learning algorithm based systems driven by increasingly available annotated clinical corpora. The representative systems include MedLEE (Friedman et al., 1994), MetaMap (Aronson and Lang, 2010), KnowledgeMap (Denny et al., 2003), cTAKES (Savova et al., 2010), etc. Clinical NLP challenges organized by the Center for Informatics for Integrating Biology & the Beside (i2b2) have promoted research using machine learning algorithms to recognize clinical entities (Uzuner et al., 2010; Uzuner et al., 2011).

Unlike the previous i2b2 challenges, the ShARe/CLEF challenge of clinical disorder extraction and encoding held in 2013 took the initiative to recognize disjoint entities, in addition to entities made up of consecutive words (Chapman et al., 2013). ShARe/CLEF challenge also required encoding of the disorder entities to Systematized Nomenclature Of Medicine Clinical Terms (SNOMED-CT) (using UMLS CUIs).

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 802–806, Dublin, Ireland, August 23-24, 2014.

In this paper, we describe our system for Task 7 of SemEval 2014, which followed the requirements of 2013 ShARe/CLEF challenge. Our system employed ensemble learning based approaches for disorder entity recognition and a Vector Space Model (VSM) based method for mapping extracted entities to CUIs of SNOMED-CT concepts. Our system was top-ranked among all participating teams according to evaluation by the organizer.

2 Method

Our end-to-end system for Task 7 of SemEval 2014 consists of two components: disorder entity recognition and encoding. The raw clinical notes first went through the pre-processing modules for rule-based sentence boundary detection and tokenization. Extracted features were then used to train two machine learning algorithm-based entity recognition models, Conditional Random Fields (CRFs) (Lafferty et al., 2001) and Structural Support Vector Machines (SSVMs) (Tsochantaridis et al., 2005), respectively. These two models were ensembled with MetaMap, a symbolic biomedical NLP system, by three different approaches. Recognized entities were mapped to SNOWMED-CT CUIs in the encoding component. Detailed information of the components are presented in the following sections.

2.1 Dataset

The training and test sets of 2013 ShARe/CLEF challenge were used as the training and development sets respectively for system development in SemEval 2014 Task 7. The training set consists of 199 notes and the development set has 99 notes, both of which were collected from four types of clinical notes including discharge summaries (DIS), radiology reports (RAD), and ECG/ECHO reports. Based on a pre-defined guideline, disorder entities were annotated for each note and then mapped to UMLS CUIs of SNOMED-CT concepts. Disorder entities not found in SNOMED-CT were marked as "CUI-less". The training set contained 5811 disorder entities which were mapped to 1007 unique CUIs or CUI-less. The development set contained 5340 disorder entities mapped to 795 CUIs or CUI-less. The test set contained 133 notes, all of which were discharge summaries. As the gold-standard annotation of the test set is not released by the organizer, the detailed annotation information of the test set is not available. Table 1 shows the total counts of notes, entities and CUIs in the three datasets.

Dataset	Type	Note	Entity	CUI	CUI-less
Train	ALL	199	5816	4177	1639
	ECHO	42	828	662	166
	RAD	42	555	392	163
	DIS	61	3589	2646	943
	ECG	54	193	103	90
Dev	ALL	99	5340	3619	1721
	ECHO	12	338	241	97
	RAD	12	162	126	36
	DIS	75	4840	3252	1588
	ECG	0	0	0	
Test	ALL	133	-	-	
	DIS	133	-	-	

Table 1. Statistics of the dataset.

2.2 Disorder entity recognition

The disorder entity recognition component consists of two modules: 1) the machine learning (e.g., CRF and SSVM) based named entity recognition (NER) module and 2) the ensemble learning module. For the challenge of this year, we mainly focused on the second ensemble learning module.

Machine learning based NER Module. This module was built based on our previous challenge participation in the 2013 ShARe/CLEF challenge (Tang et al., 2013). Annotated data were typically converted into a BIO format in machine learning-based NER systems. Each word was assigned one of the three labels: B for beginning of an entity, I for inside an entity, and O for outside of an entity. A unique challenge of this task is the high frequency (>10%) of disjoint disorders. For example, in the sentence "*the left atrium is not moderately dilated*", the discontinuous phrase "*left atrium…dilated*" is defined as a disjoint disorder. Such entities could not be directly represented using the traditional BIO approach. Therefore, in addition to traditional BIO tags used for labeling words in the consecutive disorder entities, two sets of tags were created for disjoint entities: (1) D{B, I} was used to label disjoint entity words that are not shared by multiple concepts; and (2) H{B, I} was used to label head words that belonged to more than two disjoint concepts. Ultimately, we assigned one of the seven labels {B, I, O, DB, DI, HB, HI} to each word. A few simple rules were then defined to convert labeled words to entities (Tang et al., 2013).

We exploited two state-of-the-art machine learning algorithms for disorder entity recognition, namely CRF (Lafferty et al., 2001) and SSVM (Tsochantaridis et al., 2005). CRFsuite and SVM^{hmm} were used to implement CRF and SSVM respectively.

For features, we used bag-of-word, part-of-speech from Stanford tagger, type of notes, section information, word representation from Brown clustering (Brown et al., 1992), random indexing (Lund and Burgess, 1996) and semantic categories of words based on UMLS lookup, MetaMap, and cTAKES outputs. More detailed information of this module can be found in our paper for 2013 ShARe/CLEF challenge (Tang et al., 2013).

One thing to note is that for word representation features like Brown clustering and random indexing, we only use the combination of traning and development and test datasets for feature extraction. The non-annotated corpus provided by the SemEval organizers was not employed currently. We do plan to pre-generate word clusters and random indexing using the provided corpus in the near future.

Ensemble Learning Module. Three approaches were employed to consolidate the CRF-model, SSVM-model and the MetaMap outputs, namely machine learning classifier based ensemble (ensemble^{ML}), majortiy voting based ensemble (ensemble^{MV}) and direct merging of the entity recognition results from the three models (ensemble^{DM}).

In the ensemble^{ML} approach, a binary classifier was trained to determine if the entities recognized by the CRF-model, SSVM-model and MetaMap were true positives. A new set of features were then extracted for each candidate entity, that included the specific models recognizing the entity, the entity itself, n-gram and word shape features of the first/last word of the entity. A sliding window based feature was extracted to check whether there was any recognized entity within 20 characters before the first and after the last word. Some features extracted from the first module were also employed. We used the open source toolkit Liblinear (Fan et al., 2008), to build the binary classifier for ensemble^{ML}.

2.3 Disorder Entity Encoding

We developed a Vector Space Model (VSM) based approach to find the most suitable CUI for a given disorder entity. The disorder entity was used as query and all the UMLS terms were treated as documents. We used the cosine-similarity score to rank the candidate terms. For post-processing, if the top-ranked CUI was not a disorder CUI, it was replaced with 'CUI-less'. 'CUI-less' was also assigned to entities without any retrieved candidate CUI.

2.4 Experiments and Evaluation

Our system was developed and trained using the enlarged training set by merging the 199 notes in the training set and the 99 notes in the development set. All parameters of CRF, SSVM and Liblinear were optimized by 10-fold cross-validation on the enlarged training dataset. The performance of disorder entity recognition was evaluated by precision, recall and F-measure, which were measured in both "strict" and "relaxed" modes. The "strict" mode was defined as follows: a concept is correctly recognized if and only if it can be matched exactly to a disorder mention in the gold standard, and the "relaxed" mode means that a disorder mention is correctly recognized if it overlaps with any disorder mention in the gold standard. For entity encoding, all participating systems were evaluated using accuracy, in "strict" and "relaxed" modes, as defined in (Suominen et al., 2013).

3 Results

Table 2 and Table 3 show the best performance of our systems in the SemEval 2014 Task 7 as reported by the organizers, where "P", "R", "F" denote precision, recall and F-measure respectively. For disorder entity recognition, the ensemble^{ML} based system outperformed the other two ensemble approaches, achieving the best F-measure of 0.813 under "strict" criterion and was ranked first in the challenge. For encoding, our system achieved an accuracy of 0.741 by ensemble^{DM} under "strict" criterion and was again ranked first in the challenge.

	Strict			Relaxed		
	P	R	F	P	R	F
ensemble^{ML}	84.3	78.6	81.3	93.6	86.6	90.0

Table 2. The disorder recognition performance of our system for the SemEval 2014 task 7 (%).

	Accuracy	
	Strict	Relaxed
ensemble^{DM}	0.741	0.873

Table 3. The SNOMED encoding performance of our system for the SemEval 2014 task 7.

4 Discussion

In this study, we developed an ensemble learning-based approach to recognize disorder entities and a vector space model-based method to encode disorders to UMLS CUIs. Our system was top-ranked among all participating teams. However, there are still expectations for further improvement.

For disorder entity recognition, directly merging the entity recognition results of the three models (ensemble^{DM}) achieved the highest encoding accuracy of 0.741. This shows the great potential of performance enhancement by combining different models. However, the precision of ensemble^{DM} was much lower than the current machine learning-based ensemble approach ensemble^{ML}. ensemble^{ML} improved the precision to 84.3%, with the lowest recall of 78.6% among the three ensemble approaches. Further investigations for balancing and enhancing both precision and recall simultaneously by combining different models will be pursued in the follow-up studies.

For encoding, when a disorder entity can be labelled with multiple CUIs in different contexts, a more effective disambiguation model could be exploited. Further, query expansion techniques may be helpful and worth investigating. The above methods should be potentially helpful to address the problems caused by synonyms or spelling variants.

5 Conclusion

We developed a clinical disorder recognition and encoding system that consists of a ensemble learning-based approach to recognize disorder entities and a vector space model-based method to encode the identified disorders to UMLS CUIs of SNOMED-CT concepts. The performance of our system was top-ranked in the SemEval 2014 Task 7, indicating that our approaches are promising. However, further improvements are needed in order to enhance performance on concept extraction and encoding in clinical text.

Acknowledgments

This study is supported in part by grants from NLM R01LM010681, NCI 1R01CA141307, NIGMS 1R01GM102282 and CPRIT R1307 (H.X).

Reference

Aronson, A. R., & Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association: JAMIA*, *17*(3), 229–236.

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, *32*(suppl 1), 267–270.

Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-Based n-gram Models of Natural Language. *Computational Linguistics*, *18*, 467–479.

Denny, J. C., Irani, P. R., Wehbe, F. H., Smithers, J. D., & Spickard, A. (2003). The KnowledgeMap Project: Development of a Concept-Based Medical School Curriculum Database. *AMIA Annual Symposium Proceedings*, *2003*, 195–199.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, *9*, 1871–1874.

Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J., & Johnson, S. B. (1994). A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, *1*(2), 161–174.

Lafferty, J., McCallum, A., & Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Departmental Papers (CIS)*.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association: JAMIA*, *17*(5), 507–513.

Suominen, H., SalanterÃ, S., Velupillai, S., Chapman, W. W., Savova, G., & Elhadad, N. (2013). Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, *2013*, 212–231.

Tang, B., Cao, H., Wu, Y., Jiang, M., & Xu, H. (2013). Recognizing and Encoding Discorder Concepts in Clinical Text using Machine Learning and Vector Space Model. *Workshop of ShARe/CLEF eHealth Evaluation Lab 2013*.

Tsochantaridis, I., Joachims, T., Hofmann, T., & Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, *6*, 1453–1484.

Uzuner, Ã., Solti, I., & Cadag, E. (2010). Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, *17*(5), 514–518.

Uzuner, Ã., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association: JAMIA*, *18*(5), 552–556.