TMUNSW: Disorder Concept Recognition and Normalization in Clinical Notes for SemEval-2014 Task 7

Jitendra Jonnagaddala
Translational Cancer Research Network,
University of
New South Wales,
Sydney 2031,
Australia
z3339253@unsw.edu.au

Manish Kumar
Krishagni Solutions
Pty Ltd,
Armadale 6112,
Australia
manish.kumar@krishagni.com

Hong-Jie Dai* **Enny Rachmani** **Chien-Yeh Hsu**
Graduate Institute of Biomedical Informatics,
College of Medical Science and Technology
Taipei Medical University,
Taipei City 110, Taiwan
{hjdai, d610101005, cyhsu}@tmu.edu.tw

Abstract

We present our participation in Task 7 of SemEval shared task 2014. The goal of this particular task includes the identification of disorder named entities and the mapping of each disorder to a unique Unified Medical Language System concept identifier, which were referred to as Task A and Task B respectively. We participated in both of these subtasks and used YTEX as a baseline system. We further developed a supervised linear chain Conditional Random Field model based on sets of features to predict disorder mentions. To take benefit of results from both systems we merged these results. Under strict condition our best run evaluated at 0.549 F-measure for Task A and an accuracy of 0.489 for Task B on test dataset. Based on our error analysis we conclude that recall of our system can be significantly increased by adding more features to the Conditional Random Field model and by using another type of tag representation or frame matching algorithm to deal with the disjoint entity mentions.

1 Introduction

Clinical notes are rich sources of valuable patient's information. These clinical notes are often plain text records containing important entity mentions such as clinical findings, procedures and disease mentions (Jimeno et al., 2008). Using automated tools to extract the aforementioned information can undoubtedly help researchers and clinicians with better decision making. An important subtask of information extraction called named entity recognition (NER) can recognize the boundary of named entity mention and classify it into a certain semantic group.

The focus of the SemEval-2104 task 7 is recognition and normalization of disorder entities mentioned in clinical notes. As such, this task was further divided into two parts: first, task A which includes recognition of mention of concepts that belong to UMLS (Unified Medical Language System) semantic group *disorders* (Bodenreider, 2004). The concepts considered in Task A include the following eleven UMLS semantic types: Congenital Abnormality; Acquired Abnormality; Injury or Poisoning; Pathologic Function; Disease or Syndrome; Mental or Behavioral Dysfunction; Cell or Molecular Dysfunction; Experimental Model of Disease; Anatomical Abnormality; Neoplastic Process; and Signs and Symptoms. Second, task B referred to as task of normalization involves the mapping of each disorder mention to a UMLS concept unique identifier (CUI).The mapping was limited to UMLS CUI of SNOMED clinical term codes (Spackman, Campbell, & CÃ, 1997). We participated in both tasks and devel-

*Corresponding author

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 663–667,
Dublin, Ireland, August 23-24, 2014.

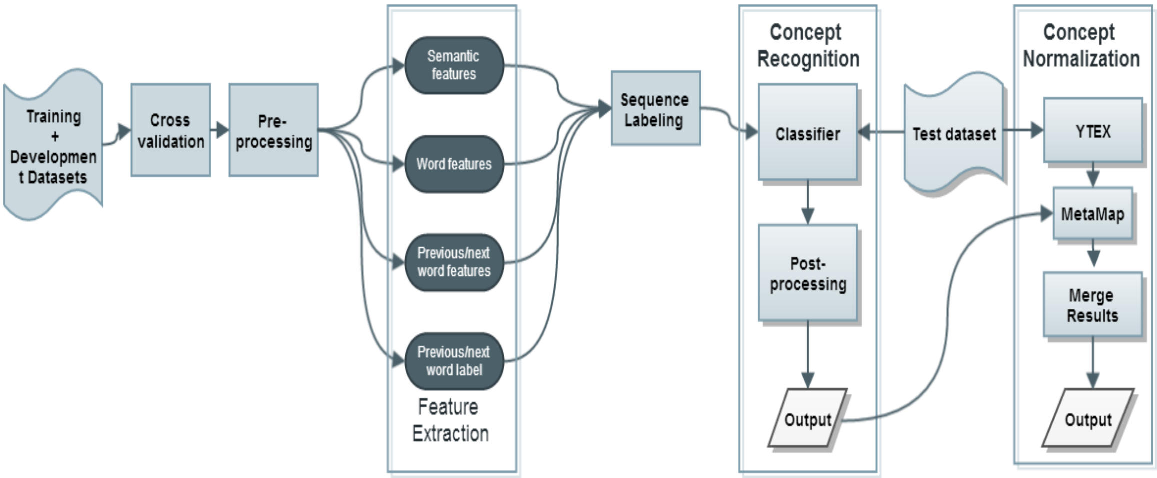

Fig. 1: TMUNSW system design for SemEval-2014 Task 7.

oped a disorder concept recognition/normalization system based on several openly available tools and machine learning algorithms.

2 Methods

2.1 System Design

For both task A and B, YTEX (Garla et al., 2011) system was employed as a baseline system. We chose to use YTEX since it is specifically designed for processing clinical notes with improvements to cTAKES's dictionary lookup algorithm and word sense disambiguation feature. The preprocessing involves sentence detection, tokenization and part-of-speech (POS) tagging(Fiscus, 1997). Based on the tokenized tokens, several features along with the corresponding part-of-speech tags were extracted for the supervised learning algorithm–conditional random field (CRF) model (Lafferty, McCallum, & Pereira, 2001). After training, the CRF model was used for recognizing disorder mentions. Furthermore the recognized disorder concepts were sent to MetaMap (Aronson & Lang, 2010) to look for their corresponding CUIs for generating normalized results. The results were finally merged with the output of YTEX. A high level diagram of the developed system is schematized in Figure 1.

2.2 Disorder Concept Recognition

The task A involves detecting boundaries of entity that belongs to UMLS semantic group, disorders. We used the sequence tagging tool based on Mallet's implementation of the supervised linear chain CRF model to perform this task. We followed the traditional BIO format to formulate the disorder concept recognition task as a sequential labelling task, wherein each token was assigned a label such as B is indicated the Beginning of entity, I is indicated the Inside an entity, or O is indicated the Outside of an entity. Thus, the model assigns each of the word into one of the above three labels. We investigated various types of features proposed in previous works (Jiang et al., 2011; Li, Kipper-Schuler, & Savova, 2008; Tang, Cao, Wu, Jiang, & Xu, 2013), like semantic feature which includes UMLS semantic group and semantic type, to develop our classifier. We also investigated various word features like POS, capitalization, and 'position of word' in the sentence. We also used 'previous word', 'next word' and 'label of these words' as a feature for developing our classifier.

2.3 Disorder Concept Normalization

Each disorder concept recognized by our recognition system was passed to a local installation of MetaMap using MetaMap Java API to obtain its candidate CUI. To increase the recall, we merged results from both YTEX and MetaMap systems. Output from YTEX baseline system was merged to the output from our CRF model with MetaMap. This method was used because it was observed that our CRF/MetaMap model has higher precision while YTEX baseline system has higher recall.

3 Results

3.1 Datasets

For Task A and Task B, the training and development datasets provided by the SemEval task 7 organizers were used. Both were derived from ShARe corpus containing de-identified plain text clinical notes from MIMIC II database (Suominen et al., 2013). These clinical notes were manually annotated for disorder mention and normalized to an UMLS CUI when possible. The corpus consisted of four types of clinical notes: discharge summaries, electrocardiogram, echocardiogram, and radiology reports. As the dataset, we included different types of clinical notes, further we trained a CRF model for each type and evaluated its performance on the corresponding development data. However, test set from task organizers contained discharge summaries only. Hence, the model developed for discharge summary was selected for evaluation on the test set.

3.2 Evaluation Metrics

The official evaluation script provided by organizers of the shared task was used to evaluate our system ability to correct an identify spans of text that belongs to semantic group disorders and to normalize them to the corresponding CUIs.

We calculated the evaluation measures under two settings-strict and relaxed. The strict setting matches exact boundaries with the gold standard, while relaxed setting matches overlapping boundaries in the gold standard. The evaluation measures were calculated using the commonly used evaluation measures including recall (R), precision (P), and F-measure (F) (Powers, 2007).

3.3 System Configurations

We used YTEX V0.8 with cTAKES V2.5.0 as the baseline system for performance comparison. All default settings for YTEX, including the concept window of the length 10, were adopted. We submit two runs for both tasks. For Task A, the first run, denoted as Run0, used the developed CRF model to recognize the disorder concepts. The second run was denoted as Run1, which merged the results of CRF model with YTEX. Similarly, for Task B, Run0 used the MetaMap 2012 version to normalize the candidate disorder concepts recognized by our CRF model. For Run1, we merged normalized annotation results of YTEX with Run0.

3.4 System Performance Comparison

We performed a ten-fold cross validation on the combination of the training and development datasets for examining the recognition and normalization performance of the developed CRF model combined with MetaMap (Run0), and compared with the YTEX as the baseline system. Table 1 summarized the results for Task A and B.

The results showed, for both tasks, Run0 significantly outperformed YTEX in the strict setting. The higher F-score of Run0 can be attributed by the fact that Run0 is developed based on the released corpus and the machine learning algorithm which is better suited for NER task as compared to the rule based YTEX system. In the relaxed setting, for Task A, Run0 also has significantly higher F-score than the YTEX baseline system. However, in case of Task B accuracy of YTEX is significantly greater than Run0. We believe that the higher accuracy of the baseline system can be attributed by the word sense disambiguation feature within YTEX.

Task A	**YTEX**		**Run0**	
	Strict	**Relaxed**	**Strict**	**Relaxed**
P	0.524	0.917	**0.771**	**0.978**
R	0.469	0.670	**0.615**	**0.811**
F	0.495	0.774	**0.682**	**0.884**
Task B	**YTEX**		**Run0**	
	Strict	**Relaxed**	**Strict**	**Relaxed**
Accuracy	0.469	**1.000**	**0.684**	0.752

Table 1. Summary of Training Set Evaluation Results.

3.5 Official Evaluation Results

Table 2 shows the official evaluation results of the submitted two configurations, Run0 and Run1. Under the strict setting, Run1 achieves the better performance with an F-measure of 0.549 for Task A and an accuracy of 0.489 for Task B on test dataset. Our best run for Task A was ranked 15 out of 21 participants, while for Task B it was ranked 9 out of 18 participants.

Task A	**Run0**		**Run1**	
	Strict	**Relaxed**	**Strict**	**Relaxed**
P	**0.622**	0.899	0.524	**0.914**
R	0.429	0.652	**0.576**	**0.765**
F	0.508	0.756	**0.549**	**0.833**
Task B	**Run0**		**Run1**	
	Strict	**Relaxed**	**Strict**	**Relaxed**
Accuracy	0.358	0.834	**0.489**	**0.849**

Table 2. Summary of Test Set Evaluation Results.

Table 2 shows that Run1 has higher F-score than Run0 because of its high recall. On the other hand, Run0 achieves significantly higher precision compared to Run1 for Task A. The result is in accordance with our expectation, because Run1 integrated the results from YTEX to improve the recall of Run0 at the cost of the decrease in precision. The trade-off seems acceptable because it can significantly improve the accuracy in normalizing disorder concepts.

4 Discussion

We performed error analysis on development dataset and found that the lower recall of Run0 derived from the miss of many disjoint entities (where the tokens comprising the entity string are non-adjacent), which cannot be captured by the current BIO tag set. For example, consider the sentence "*Abdomen is soft, nontender, nondistended, negative bruits.*" For this sentence the gold annotations contain three entities as "*Abdomen bruits*-CUI= C0221755", "*Abdomen nontender*-CUI=CUI-less" and "*nondistended*-CUI=CUI-less". In the current BIO formulation, all of the above three disjoint entities cannot be correctly recognized. There are also abbreviations which were rarely seen in the training dataset but appeared more in the development/test sets. So when we test our developed model on test set the abbreviations which are not part of training and development set must have been missed by our system. We believe that by incorporating medical abbreviations database into our model development, the performance of our overall system would have been better. Also, the precision in Task A of Run1 was lower than Run0 because of some disjoint annotations.

5 Conclusion

We present a clinical NER system based on Mallet's implementation of CRF and a hybrid normalization system using MetaMap and YTEX. We developed our system with limited features due to the time constraint. We can conclude from error analysis that recall of this system could be significantly increased by adding more features to it. We plan to extend our system in future by using another type of tag representation or frame-based pattern matching algorithm to handle disjoint named entities. Similarly missing abbreviations can be handled by employing external resources such as abbreviation recognition tools.

Acknowledgements

This project was supported by a Cancer Institute NSW's translational cancer research centre program grant, and the research grant TMU101-AE1-B55 of Taipei Medical University.

References

Aronson, A. R., & Lang, F. M. (2010). An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc, 17*(3), 229-236. doi: 10.1136/jamia.2009.002733

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research, 32*(suppl 1), D267-D270.

Fiscus, J. G. (1997). *A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER).* Paper presented at the Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on.

Garla, V., Re, V. L., Dorey-Stein, Z., Kidwai, F., Scotch, M., Womack, J., . . . Brandt, C. (2011). The Yale cTAKES extensions for document classification: architecture and application. *Journal of the American Medical Informatics Association, 18*(5), 614-620.

Jiang, M., Chen, Y., Liu, M., Rosenbloom, S. T., Mani, S., Denny, J. C., & Xu, H. (2011). A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association, 18*(5), 601-606.

Jimeno, A., Jimenez-Ruiz, E., Lee, V., Gaudan, S., Berlanga, R., & Rebholz-Schuhmann, D. (2008). Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics, 9 Suppl 3,* S3. doi: 10.1186/1471-2105-9-S3-S3

Lafferty, J., McCallum, A., & Pereira, F. (2001). *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. Paper presented at the Proceedings of the 18th International Conference on Machine Learning (ICML).

Li, D., Kipper-Schuler, K., & Savova, G. (2008). *Conditional random fields and support vector machines for disorder named entity recognition in clinical texts.* Paper presented at the Proceedings of the workshop on current trends in biomedical natural language processing.

Powers, D. M. W. (2007). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies, 2*(1), 37-63. doi: citeulike-article-id:10061513

Spackman, K. A., Campbell, K. E., & CÃ, R. (1997). *SNOMED RT: a reference terminology for health care.* Paper presented at the Proceedings of the AMIA annual fall symposium.

Suominen, H., Salanterä, S., Velupillai, S., Chapman, W. W., Savova, G., Elhadad, N., . . . Jones, G. J. (2013). Overview of the ShARe/CLEF eHealth Evaluation Lab 2013 *Information Access Evaluation. Multilinguality, Multimodality, and Visualization* (pp. 212-231): Springer.

Tang, B., Cao, H., Wu, Y., Jiang, M., & Xu, H. (2013). Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. *BMC medical informatics and decision making, 13*(1), 1-10.