Combining MetaMap and cTAKES in Disorder Recognition: THCIB at CLEF eHealth Lab 2013 Task 1

Yunqing Xia[1], Xiaoshi Zhong[1]
Peng Liu[2], Cheng Tan[2], Sen Na[2], Qinan Hu[2], Yaohai Huang[2]

[1] Dept. of Comp. Sci. & Tech., Tsinghua National Laboratary of Information Science and Technologies, Tsinghua University, Beijing 100084, China
{yqxia, xszhong}@tsinghua.edu.cn

[2] Canon Information Technology (Beijing) Co. Ltd., Beijing 100080, China
{liupeng, tancheng, nasen, huqinan, huangyaohai}@canon-ib.com.cn

Abstract. This paper describes the THCIB systems that used in the ShARe/CLEF eHealth 2013 task 1. We implemented two baseline systems and a combination system using the existing technologies. One baseline system is built using MetaMap. We built another baseline system using cTAKES. Furthermore, we developed the combination system with a system combination method. The results of combination system were submitted because the combined results performed better than either single system. We also report the experimental results on the training set and the test set.

Keywords: disorder recognition, disorder normalization, clinical report processing, natural language processing, information extraction

1 Introduction

The ShARe/CLEF eHealth Lab 2013 task 1 aims at named entity recognition and normalization of disorders [1]. There are two subtasks: 1a) recognition of mentions of concepts that belong to UMLS semantic group disorders; and 1b) mapping each mention to a unique UMLS CUI (Concept Unique Identifier). For example, an input sentence is "The rhythm appears to be atrial fibrillation". Task 1a aims to recognize disorder "atrial fibrillation", and task 1b aims to map the disorder to CUI "C0004238". This year we participated in both subtasks.

For the time limitation, we built the baseline systems and combination system using existing technologies. The results of combination system were submitted due to better performance. In this paper we describe the workflow of the baseline systems and combination system. And we also present the experimental results on the training set and the test set.

The reminder of this paper is structured as follows. In section 2, we present an overview of our baseline systems. In section3, we describe the system combination method. The experiments and analysis of the result are described in section 4. We give the conclusion in section 5.

2 Baseline Systems

We built two baseline systems for task 1. Both baseline systems are implemented using open source software (OSS). One is built using MetaMap [2], and the other is built using cTAKES [3].

2.1 Baseline System 1: MetaMap

MetaMap is a highly configurable program developed by the National Library of Medicine (NLM) to map biomedical text to the UMLS (Unified Medical Language System) [4] Metathesaurus or, equivalently, to discover Metathesaurus concepts referred to in text [5].

The flowchart of baseline system 1 is shown in Fig. 1. In the baseline system 1, the clinical text is processed as following steps: 1) the clinical text is sent to MetaMap; 2) the MetaMap processes the clinical text and maps all Metathesaurus concepts in the clinical text to UMLS. The concepts will be saved in an XML file. 3) Post-processing the XML file, and extract the disorders and the corresponding CUIs; 4) output the disorders and CUIs.

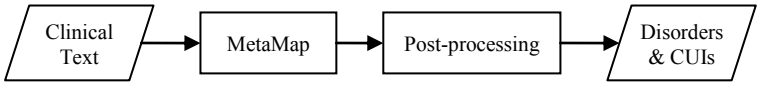

Fig. 1. Flowchart of baseline system 1.

2.2 Baseline System 2: cTAKES

cTAKES (Apache clinical Text Analysis and Knowledge Extraction System) is an open source natural language processing system for information extraction from electronic medical record clinical free-text [6]. It can process the clinical text and identify the clinical named entities from various dictionaries including the UMLS. Each entity has attributes such as the text span, the ontology mapping code, etc..

The flowchart of baseline system 2 is shown in Fig. 2. The clinical text is processed as following steps: 1) the clinical text is sent to cTAKES; 2) the cTAKES processes the clinical text, and extract named entities. The extracted named entities will be stored in an XCAS file; 3) post-processing the XCAS file, and extract the named entities which belong to disorders and the corresponding CUIs; 4) output the disorders and CUIs.

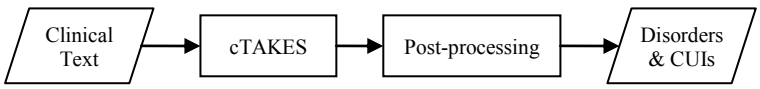

Fig. 2. Flowchart of baseline system 2.

3 Combination System: The Submitted Run

In order to take advantage of the strengths of two baseline systems, we employ a simple but efficient system combination approach to combine the results of the baseline systems [7]. We express the results of baseline system 1 as X, where $X = \{x_1, x_2, \ldots, x_m\}$; and we express the results of baseline system 2 as Y, where $Y = \{y_1, y_2, \ldots, y_n\}$. And we express the results of combination system as Z. Then the combination algorithm is shown as follows.

```
ALGORITHM 1:
Set Z empty; //
for (i=1; i<=m; i++) { // Initialization
    add x_i to Z; //
}

for (j=1; j<=n; j++) { // Combination
    if y_j conflict with Z:
        discard y_j;
    else:
        add y_j to Z;
}
```

This is because that the baseline system 1 has higher precision, while the baseline system 2 has higher recall.

4 Experimental Results

4.1 Dataset

The training set contains 200 clinical reports, and totally 5874 disorders. We used all of the training set to evaluate the performance of each system. The test set contains 100 clinical reports. We will give evaluation results on both training set and test set.

4.2 Evaluation Metrics

Precision, recall and F1 measure are used in this evaluation. Two conditions are setup. One is strict, which means that the recognized words are perfectly matched; the other is relaxed, which means that the recognized words have overlap with the gold standard.

4.3 Internal Results

We evaluate three systems using training set. The results of task 1a and task 1b are shown in Table 1 and Table 2, respectively. The baseline1 is the results of baseline system 1; the baseline2 is the results of baseline system 2; and the combination is the results of the combination system.

Table 1. Evaluation results of Task1a on the training set.

Task 1a	*Baseline1*		*Baseline2*		*Combination*	
	Strict	Relaxed	Strict	Relaxed	Strict	Relaxed
Precision	0.636	0.789	0.415	0.763	0.413	0.733
Recall	0.463	0.573	0.428	0.641	0.521	0.725
F-score	0.536	0.664	0.422	0.697	0.461	0.729

Table 2. Accuracy of Task1b on the training set.

Task 1b	*Baseline1*	*Baseline2*	*Combination*
Strict	0.401	0.389	0.455
Relaxed	0.866	0.910	0.873

Table 3. Task1a evaluation results on the test set.

Task 1a	*test set*	
	Strict	Relaxed
Precision	0.445	0.720
Recall	0.551	0.713
F-score	0.492	0.716

Table 4. Task1b evaluation results on the test set.

Task 1b	*test set*	
	Strict	Relaxed
Accuracy	0.470	0.853

From Table 1, we can find that the baseline1 performs better in the strict metric while baseline2 performs better in the relaxed metric. According to analysis of the results, we find that the average length of baseline2 results is shorter than baseline1 results, but the quantity is larger than the baseline1 results. This leads to a higher recall but lower precision.

From the results of combination, we find that the recall get great improvement. And the F-score can also be improved, especially for the relaxed metric.

Table 2 shows the evaluation results of task 1b. The baseline1performs better in the strict metric and baseline2 performs better in the relaxed metric. After combining the results of two baseline system, the results can improve from 0.401 to 0.455 in the strict metric and acceptable decrease in the relaxed metric.

4.4 Official Results

Table 3 and Table 4 show the official evaluation results of combination system on the test set. The combination system is robust because it has similar performance on the training set and test set.

5 Conclusion

For the time limitation, our main purpose is using the existing technologies to build baseline system for disorder recognition and verify the performance of the existing technologies. We built two baseline systems using OSS for ShARe/CLEF eHealth task 1. And we also built a combination system by combine the results of two baseline systems. The evaluation results on the training set and the test set show that the combination system can perform better than single baseline system.

Acknowledgement

This research is supported by Canon Inc. (No. QIM2013). The Shared Annotated Resources (ShARe) project is funded by the United States National Institutes of Health with grant number R01GM090187. We also appreciate the valuable comments from the task organizer.

References

1. Suominen, Hanna, Sanna Salanter, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Danielle Mowery, Johannes Leveling, Lorraine Goeuriot, Liadh Kelly, David Martinez and Guido Zuccon. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. Proceedings of CLEF 2013. Lecture Notes in Computer Science (LNCS), Springer.
2. MetaMap, http://mmtx.nlm.nih.gov/
3. Apache cTAKES, http://ctakes.apache.org/index.html
4. Unified Medical Language System (UMLS), http://www.nlm.nih.gov/research/umls/
5. Aronson, A.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Journal of the American Medical Informatics Association, pp. 17—21 (2001)
6. Savova, G., Masanz, J.J., Ogren, P., Zheng, J., Sohn, S., Kipper-Schuler, K., Chute, C.: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component, evaluation and applications. J Am Med Inform Assoc. 17, 507-513 (2010)
7. Fiscus, J.G.: A post-processing system to/yield reduced word error rate: Recognizer Output Voting Error Reduction (ROVER). In Proc. IEEE workshop on automatic speech recognition and understanding. (1997)