Disease Template Filling using the CTAKES YTEX Branch

John David Osborne

University of Alabama at Birmingham, Birmingham AL 35294, USA,
ozborn@uab.edu,
WWW home page: http://coral.cis.uab.edu/

Abstract. Using an adapted version of the YTEX branch of CTAKES for disease template filling accuracies of 0.936, 0.974, 0.807 and 0.926 were achieved for the conditional, generic, negation and subject class respectively in Task 2a. Overall accuracy was 0.79. Unfortunately substantially poorer performance in F1 score, precision and recall for all 4 of these templating tasks indicates that it is not yet possible to get good performance using these CTAKES algorithms in this task.

Keywords: CTAKES, YTEX, evaluation, information extraction

1 Approach and Objectives

The YTEX [1] development branch of CTAKES [2] pipeline was evaluated for template filling (Task 2a) [3]. The objective was to use the existing CTAKES tools to populate the template for negation, subject class, conditional qualifiers and generic references and to use the YTEX word sense disambugation and dictionary lookup component to identify the anatomic location of the disease. The remaining template filling tasks were not attempted and the YTEX based anatomical location lookup was not completed in time for the test data.

2 Methodology

The base system employed was the YTEX branch of ctakes, specifically revision 1588688 at https://svn.apache.org/repos/asf/ctakes/branches/ytex. Default settings were used for YTEX, including a concept window length of 10. The 2013AB version of UMLS was used. Identified annotations matching the appropriate disease UMLS semantic types were checked for overlap with input disease templates as defined in the Share schema [4]. The CTAKES generated modifiers were then used to fill the template, otherwise the default values were used to fill the template. No machine learning or training on the provided data took place.

The system also included some additional non-CTAKES rule-based annotators from a previous system [5] designed for ShARe/CLEF eHealth 2013 concept recognition. However the only role they played was to better match CTAKES

generated identified annotations to ShARe/CLEF eHealth 2014 disease concepts; not to fill out the disease templates. Additionally the system also included an annotator capable of recognizing a variety of different section types in clinical notes. This annotator was developed on a variety clinical notes at the University of Alabama at Birmingham (UAB) including discharge summaries and was not otherwise modified in time for the test data. It was employed here only to find family history sections in clinical notes and to change the subject to family for disease occurrences in this section.

3 Results

Table 1. CORAL System Task 2a Test Results

Task	Rank	Accuracy	F1 Score	Precision	Recall
Overall average	10	0.790	0.030	0.240	0.016
Norm BL	8	0.546	0	0	0
Norm CC	4	0.961	0	0	0
Norm CO	5	0.936	0.052	0.500	0.028
Norm DT	9	0.001	0	0	0
Norm GC	3	0.974	0	0	0
Norm NI	12	0.807	0.196	0.746	0.113
Norm SC	8	0.926	0.161	0.098	0.450
Norm SV	6	0.942	0	0	0
Norm TE	1	0.864	0	0	0
Norm UI	3	0.941	0	0	0

All template tasks with an F1 Score, precision and recall of zero were not attempted by the CORAL system with the exception of generic mentions (Norm GC). In the case of generic mentions, the CTAKES based generic determination did not identify any in the test data although it was actively searching for them. In the Norm SC (Subject Class) task, the use of UAB family history section identification was not useful, the regular expressions developed for identifying family history for UAB notes were not triggered on the test data. This underscores the diversity of clinical notes and the frailty of regular expression based approaches. Finally, individual results for other tasks indicate that it is possible to achieve seemingly reasonable accuracy in this task just by filling in the default value for the template.

4 Analysis and Discussion

The overall poor performance of the CTAKES based template filling for the 4 attempted tasks indicates that no off the shelf solution exists for this type of disease concept templating.

Acknowledgements This project was supported by the UAB Center for Clinical and Translational Science - grant number UL1 RR025777 from the NIH National Center for Research Resources, and the UAB Office of the Vice President for Information Technology.

References

1. Garla, V., Re III, V.L., Dorey-Stein, Z., Kidwai, F., Scotch, M., Womack, J., Justice, A., Brandt, C.: The Yale cTAKES extensions for document classification: architecture and application. J. Am. Med. Inform. Ass. 18 614–620 (2011)
2. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J. Am. Med. Inform. Ass. 17 507–513 (2010)
3. L Kelly, L Goeuriot, H Suominen, T Schreck, G Leroy, DL Mowery, S Velupillai, WW Chapman, D Martinez, G Zuccon, J Palotti: Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. Springer-Verlag.
4. N Elhadad, W Chapman, T O'Gorman, M Palmer, G Savova. The ShARe Schema for the Syntactic and Semantic Annotation of Clinical Texts. In preparation.
5. Osborne, J. D., Gyawali, B., Solorio, T.: Evaluation of YTEX and MetaMap for clinical concept recognition. arXiv preprint arXiv:1402.1668 (2014)