Integrated cTAKES for Concept Mention Detection and Normalization

Hongfang Liu, Kavishwar Wagholikar, Siddhartha Jonnalagadda, Sunghwan Sohn

Mayo Natural Language Processing Program, Mayo Clinic
200 First Street SW, Rochester, MN 55905, USA

Abstract. We participated Task 1 using an existing system MedTagger implemented in integrated cTAKES (icTAKES). The concept mention detection is based on Conditional Random Fields (CRF) and the concept mention normalization is based on a greedy dictionary lookup algorithm. A distinctive feature in MedTagger compared to other concept mention detection systems is the incorporation of dictionary lookup results into a machine learning framework for sequential labeling. Dictionary lookup results of MedLex and semantic vectors representing distributed semantics were used as features. Overall, the precision, recall, and F-measure of our best run for concept mention are 0.8, 0.573, and 0.668 respectively for strict evaluation and 0.939, 0.766, and 0.844 for relaxed evaluation. The accuracy of our best run for concept mention normalization is 54.6% and 87.0% for strict and relaxed mapping, respectively.

Keywords: named entity recognition, dictionary lookup, normalization, conditional random fields, distributed semantics

1 Introduction

Concept identification from free text is a critical component in natural language processing (NLP) applications that extract clinical or biomedical information from free text. Concept identification can be split into two steps. The first step, concept mention detection, involves the detection of text spans containing concepts of interest. And the second step, concept mention normalization, maps text spans detected to concept identifiers present in standard terminologies or ontologies. In NLP share-task workshops such as BioCreAtive or I2B2 NLP workshops [1-3], sequential labeling algorithms (i.e., Conditional Random Fields (CRF)) and machine learning methods (i.e., Support Vector machine (SVM)) have been demonstrated to achieve promising performance when provided with a large annotated corpus for training. The availability of machine learning software packages, such as SVMstruct, YamCha, MALLET, and CRFSuite, has boosted the baseline performance of concept mention detection systems. Concept mention normalization has not been tackled and the normalization tasks defined in the NLP challenge workshops were to assign gene/protein identifiers to abstracts [2] or diagnosis to a patient [4], not individual mentions in text.

In the past, we participated gene/protein name tagging and normalization tasks in BioCreAtive workshops [5,6] and developed a tagging system called BioTagger-GM [7]. We then adapted BioTagger-GM to MedTagger for clinical concept mention detection in I2B2 NLP Challenge 2010 and 2012 [8,9]. Recently, we incorporated MedTagger into

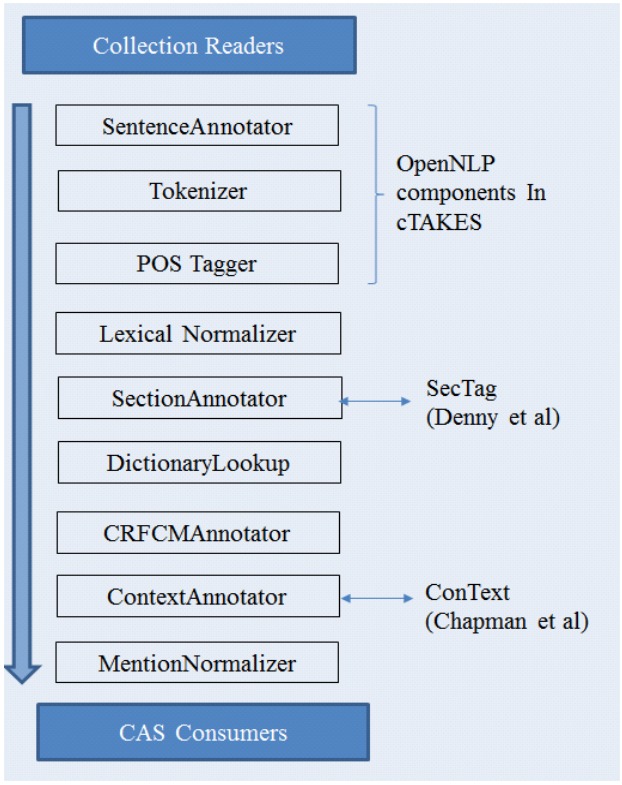

Figure 1. Component annotators in icTAKES

integrated cTAKES (icTAKES). For the SHARE/CLEF NLP Task 1, we used the icTAKES version of MedTagger.

2 System Description

Figure 1 shows the component annotators in MedTagger. We use the cTAKES wrappers of openNLP components for sentence detection, tokenization, and part-of-speech parsing and incorporate a rule-based section tagger based on SecTag and a rule-based context annotator based on ConText. The concept mention detection is based on machine learning and the concept mention normalization is based on a greedy algorithm. A distinctive feature in MedTagger compared to other concept mention detection systems is the incorporation of dictionary lookup results into a machine learning framework for sequential labeling. In the following, we describe the details of the dictionary lookup and machine learning.

2.1 Dictionary Lookup

The dictionary lookup approach in MedTagger uses the Aho-Corasick string matching algorithm. In the lookup lexical variants, punctuations, and stop words are ignored. Given a dictionary, the alphabetic set in the algorithm consists of all tokens in the dictionary. Figure 2 illustrates the representation of four dictionary entries ("GI Bleed", "acute GI bleed", "acute pain", "bleed") as a tree in the Aho-Corasik algorithm. MedTagger allows three different ways of dictionary lookup: exact string matching, lower case string matching, and flexible string matching. An example of flexible string matching is provided in Figure 2. In flexible string matching, stop words and punctuation marks are ignored and lexical variants are normalized to their base form using the Specialist Lexicon. MedTagger gives the option of returning all possible matches or the longest matches from left to right.

2.1 Mention Detection

When provided an annotated corpus, MedTagger uses CRF to detect concept mentions. For a given tokenized document, concept mention detection can be treated as a sequential labeling task where each token (e.g., word) is labeled with an appropriate label (B, I, and O) to demarcate concept mentions. Here, the label B indicates the token is the beginning of a concept mention, I the middle of a concept mention, and O the tokens not part of a concept mention. Each token is represented by features, which include the token itself as one type of features. Besides widely-used features, such as nearby words and suffixes within a window size, MedTagger incorporates dictionary lookup results as features (see [7] for details). If a phrase in the text (sequence of tokens) is mapped to a dictionary entry, the phrase is assigned with labels "*L_SemT*", where *L* is one of the three labels, B, I and O, and *SemT* is the type of the phrase in the designated dictionary, e.g., UMLS semantic type. Note that it is possible that multiple labels are assigned in case of overlapping mapping.

2.2 Mention Normalization

Both dictionary entries and detected mentions can be compositional at different granularities. For example, "enlarged spleen" can appear in text as "spleen is enlarged". Or there is no concept as "enlarged" in the dictionary but in the text, we have "enlarged spleen". There are two steps in mention normalization. The first step is to find the minimum number of dictionary entries corresponding to the detected mentions. The detail approach is described in a previous paper [10]. The second step is to search for mappings that span multiple spans. Basically, all dictionary entries are processed to capture the compositional structure. Spans located near each other are then composed to see the possibility of mapping to a dictionary entry.

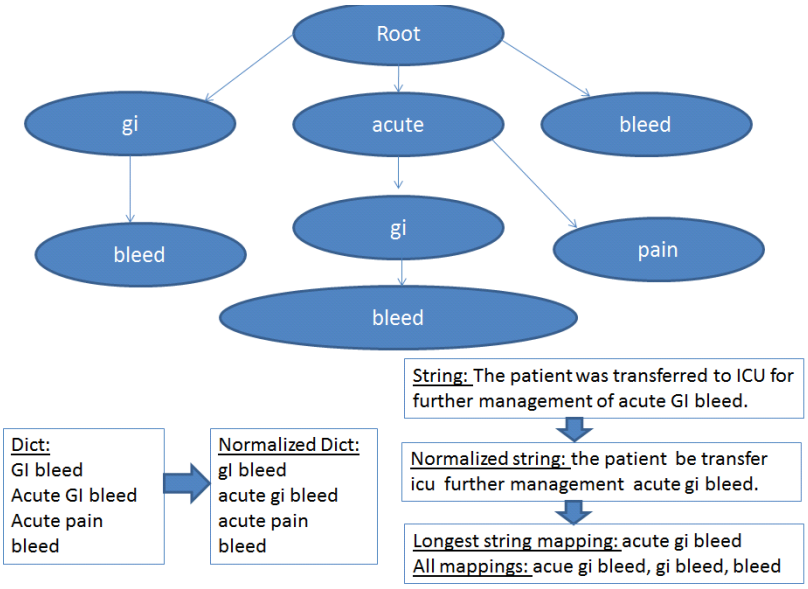

Figure 2. An illustration of MedTagger dictionary lookup.

3 Submission Description

The training set consisted of de-identified 200 clinical reports with standoff annotations of disorder mention spans and UMLS concept unique identifiers (CUIs) and test set had 100 clinical reports.

In addition to features deployed in MedTagger, for this challenge, we implemented automatically generated distributional semantic features based on a semantic vector space model trained from unannotated corpora from Mayo Clinic's clinical notes and MIMIC dataset. This model, referred to as the directional model, uses a sliding window that is moved through the text corpus to generate a reduced-dimensional approximation of a token-token matrix, such that two terms that occur in the context of similar sets of surrounding terms will have similar vector representations after training. The semantic vector for a token is obtained by adding the contextual vectors gained at each occurrence of the token, which are derived from the index vectors for the other terms it occurs within the sliding window. The model was built using the open source Semantic Vectors package [11]. Previous experiments [12,13] revealed that using directional model with 2000-dimensional vectors, five seeds (number of +1s and –1s in the vector), and a window radius of six is better suited for the task of named entity recognition. While a stop-word list is not employed, we have rejected tokens that appear only once in the unlabeled corpus or have more than three non-alphabetical characters. Note that the dictionary used here is MedLex [14].

CRFSuite was used with default setting to train first order CRF models on the training datasets. We limited the training set to ECHO, RADIOLOGY and DISCHARGE notes. A window of two tokens to the left and one token to the right was used to aggregate features for each token. To evaluate the effectiveness of features we measured system performance by excluding one feature type at a time. Table 1 shows the listing of the features, in decreasing order of their effectiveness for system performance. The final submissions were based on all features. We did not apply post processing rules.

The default output from MedTagger gene mention was submitted as Run 2 for Task 1a and Run 1 for Task 1a was obtained by supplementing Run 2 with multi-spans appearing in the training data. We submitted two runs for Task 1b (mention normalization) where Run 1 was based on concept mentions detected in Task 1a Run1 and Run 2 was supplementing with spans detected using dictionary lookup. We limited both runs to only SNOMED CT CUIs. In case of ambiguity, we sorted all CUIs in ascending order and used the first one.

4 Results and Discussion

For Task 1a our system ranked fourth and third in the strict and relaxed evaluation, respectively (Table 2). The precision of our system was equal to the best system for strict evaluation but exceeded the best system in the relaxed evaluation. In Task 1b our system ranked second and third for the strict and relaxed evaluations, respectively (Table 3).

Table 1. Performance of the system after excluding a particular feature

Feature	Strict F1	Relaxed F1
Semantic Group	0.642	0.850
Capitalization	0.649	0.860
UMLS Preferred name	0.652	0.842
Use of BI notation for semantic group	0.654	0.862
Suffixes	0.662	0.854
Sematic vector from Mayo data	0.663	0.859
Normalized form	0.666	0.849
Certainty	0.670	0.847
Section header	0.671	0.854
Part of speech	0.671	0.858
Prefixes	0.673	0.856
Punctuation	0.674	0.869
Semantic vector from MIMIC	0.674	0.862
Without excluding any feature	0.678	0.861

Table 2. Relative Performance for Task 1a

Strict			
System	**Precision**	**Recall**	**F-score**
Best team	0.800	0.706	0.750
TeamMayo (Rank 4)	0.800	0.573	0.668
Average	0.603	0.478	0.513
Relaxed			
Best team	0.925	0.827	0.873
TeamMayo (Rank 3)	0.939	0.766	0.844
Average	0.807	0.650	0.686

Table 3. Relative Performance for Task 1b

Strict	
System	**Accuracy**
Best team	0.589
TeamMayo (Rank 2)	0.546
Average	0.362
Relaxed	
Best team	0.939
TeamMayo (Rank 3)	0.870
Average	0.652

Our participation in the NLP challenge provides us with valuable knowledge in further performance improvement of concept mention and normalization, especially concept normalization. Note that we purposely did not perform rigorous training based on the training data as well as deploying post processing rules due to the assumption that tuning a system according to a specific annotated corpus too much may not perform well for a different set of samples annotated by a different research team [15,16].

Acknowledgement

The work was supported by ABI: 0845523 from United States National Science Foundation, R01LM009959 from United States National Institute of Health. The challenge was organized by the Shared Annotated Resources (ShARe) project, funded by the United States National Institutes of Health with grant number R01GM090187.

5 References

1. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc.* Sep-Oct 2011;18(5):552-556.
2. Morgan AA, Lu Z, Wang X, et al. Overview of BioCreative II gene normalization. *Genome Biol.* 2008;9 Suppl 2:S3.
3. Smith L, Tanabe LK, Ando RJ, et al. Overview of BioCreative II gene mention recognition. *Genome Biol.* 2008;9 Suppl 2:S2.
4. Uzuner O. Second i2b2 workshop on natural language processing challenges for clinical records. *AMIA Annu Symp Proc.* 2008:1252-1253.
5. Liu H, Torii M, Hu Z, Wu CH. Gene Mention and Gene Normalization Based on Machine Learning and Online Resources. Paper presented at: Proceeding of BioCreAtIve II workshop2007.
6. Hirschman L, Colosimo M, Morgan A, Yeh A. Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics.* 2005;6 Suppl 1:S11.
7. Torii M, Hu Z, Wu CH, Liu H. BioTagger-GM: a gene/protein name recognition system. *J Am Med Inform Assoc.* Mar-Apr 2009;16(2):247-255.
8. Torii M, Wagholikar K, Liu H. Using machine learning for concept extraction on clinical documents from multiple data sources. *J Am Med Inform Assoc.* Sep-Oct 2011;18(5):580-587.
9. Sohn S, Wagholikar KB, Li D, et al. Comprehensive temporal information detection from clinical text: medical events, time, and TLINK identification. *J Am Med Inform Assoc.* Apr 4 2013.
10. Liu H, Wagholikar K, Wu ST. Using SNOMED-CT to encode summary level data - a corpus analysis. *AMIA Summits Transl Sci Proc.* 2012;2012:30-37.
11. Widdows D, Cohen T. The Semantic Vectors Package: New Algorithms and Public Tools for Distributional Semantics. *Fourth IEEE International Conference on Semantic Computing.* Vol 12010.
12. Jonnalagadda S, Cohen T, Wu S, Gonzalez G. Enhancing clinical concept extraction with distributional semantics. *J Biomed Inform.* Feb 2012;45(1):129-140.
13. Jonnalagadda S, Cohen T, Wu S, Liu H, Gonzalez G. Using Empirically Constructed Lexical Resources for Named Entity Recognition. *Biomed Inform Insights.* 2013.
14. Liu H WS, Li D, Jonnalagadda S, Sohn S, Wagholikar K, Haug PJ, Huff SM, Chute CG Towards a semantic lexicon for clinical natural language processing Paper presented at: Annual Symposium of American Medical Informatics Association2012; Chicago.
15. Wagholikar KB, Torii M, Jonnalagadda SR, Liu H. Pooling annotated corpora for clinical concept extraction. *Journal of Biomedical Semantics.* 2013;4(1):3.
16. Wagholikar K, Torii M, Jonnalagadda S, Liu H. Feasibility of pooling annotated corpora for clinical concept extraction. *AMIA Summits Transl Sci Proc.* 2012;2012:38.