Performance of a multi-class biomedical tagger on clinical records

S. V. Ramanan[1], Shereen Broido[2], and P. Senthil Nathan[1]

[1] RelAgent Pvt Ltd, Chennai, India,
ramanan,senthil@npjoint.com, http://npjoint.com
[2] Vidhodaya Schools, Chennai, India,
shereen.broido@gmail.com

Abstract. We tested the performance of Cocoa, an existing dictionary/rule based entity tagger that tags multiple semantic types in biomedical domain including diseases, on disease/sign/symptom detection in clinical records in the ShARe/CLEF eHealth task. Initial analysis showed that the precision was high ($\geq 90\%$), but recall was low ($\approx 50\%$) due to (a) phrases peculiar to clinical notes (b) disambiguation of common words and (c) the large number of undefined acronyms. We extended the system to handle these cases by reference to the local intrasentential context as derived from the training set. A small module was also added for event-based detection of annotated sentence fragments containing verbs/gerunds; an example is 'LV systolic function appears depressed'. The event detection system had about 30 rules. With these modifications, the f-score was 0.75 on the test set. In a second run, we added about 70 frequently occurring acronyms as well 15 phrases which were all in caps. The final results on the test set ($f = 0.78$) show that a multi-class tagger can work reasonably well on clinical records.

Keywords: rule-based tagger, multiple entity types, clinical notes.

1 Background

Automatically tagging and normalizing mentions of diseases, signs and symptoms in clinical records is a useful addition even when these records have already been manually assigned ICD codes. Automatically assigned tags may help uncover unexpected correlations between symptoms [5], and may also be useful in checking the accuracy of the manual annotations.

Previous shared tasks in the clinical domain have addressed subsets of records that are typically seen by a medical practitioner, such as radiology reports [3] and discharge summaries [7]. The current ShARe/CLEF eHealth task covers annotation of diseases, signs and symptoms in a mixed bag of documents, including discharge summaries and echo, radiology and ECG reports [6]. However, the task does not cover GP notes, a very challenging category [1].

Cocoa [4] is an existing named entity tagger for published literature in the biomedical domain. Cocoa tags entities across a variety of semantic classes, including chemicals, proteins, cellular parts, anatomical parts and diseases. We wished to explore how well such a system would perform on clinical notes, which is a domain slightly different in scope and context from published biomedical literature. For example, common terms and phrases, such as ‘mass’ and ‘effusion’, refer exclusively to signs/symptoms when used in clinical notes. We explored whether sentence-level disambiguation is sufficient to resolve such ambiguous phrases. Additionally, acronyms are well-understood in the clinical context, and therefore used without an associated expansion in discharge summaries for example. With a small subset of pre-defined long acronyms, context-sensitive short acronyms, and with sentence-level disambiguation, the system gave a precision of 0.90 and a recall of 0.69 for a f-score of 0.78. While much less than the top-ranked score ($f = 0.87$), the results show that multi-class recognition systems can produce reasonable performance in the clinical domain.

2 System pipeline

A schematic of the system is given in Figure 1. As mentioned above, the system tags entities across a number of semantic classes. We restrict our discussion to entities relevant to the present task, namely diseases, signs and symptoms, along with the anatomical parts that they affect.

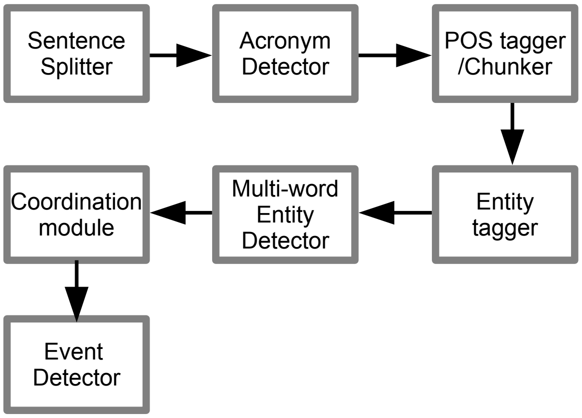

Fig. 1. Block-level pipeline of modules in the Cocoa NER system

(a) Sentence splitter. In the biomedical literature, sentence boundaries present a challenge as sentences can span multiple lines. Further, sentences can begin with lower case letters (e.g. ‘cAMP’) and numbers, as many biological entities have complex, but widely recognized, orthography. However, we observed in the CLEF training set that, both in clinical notes and in lab reports, sentences often did not have a trailing period (‘full stop’). Moreover, sentence fragments were often used to describe mental states or conditions for example. We therefore used newlines to mark sentence boundaries.

(b) Acronym detector. The system detects acronyms through a dynamic programming methodology. Even though we did not note any acronym definitions in the training set, this module was not disabled.

(c) POS tagging and chunking. These were done by TBL methods, with a Brill POS tagger followed by a fnTBL-based chunker. Both are already heavily modified for the biomedical domain in the existing system, and we did not make any substantial changes for this task.

(d) Entity tagging. Both anatomical parts and diseases were tagged at a word level with the help of word-level dictionaries (e.g., ‘Parkinsonism’) and dictionaries of morphological prefixes and suffixes (e.g. ‘cephalon’ for anatomy and ‘oglossia’ for diseases). False positive dictionaries were maintained for entities detected by morpheme-based methods. Tags were used in a limited way to correct chunking errors, primarily in VP chunks.

(e) Multi-word entities. Adjectives such as ‘aberrant’, ‘ruptured’ and ‘enlarged’ followed by an anatomical part are tagged as a symptom. We also tagged other adjectives connected with time (e.g. ‘postictal’) preceding diseases, disease postpositions (‘progressiva’), as well as a host of domain-dependent word combinations (‘prominent ear’, ‘wasting disease’). These multi-word combinations were derived from an exhaustive analysis of UMLS and ICL definitions of diseases, signs and symptoms. Multi-word combinations involving anatomical parts were derived from a number of sources, including Gray’s Anatomy. For the CLEF task, we extended this module to disambiguate common words as signs/symptoms with appropriate context (‘negative drift’, ‘negative masses’, ‘bilateral effusion’, ‘adventitious movements’). A narrow context/trigger based tagging was also added for certain acronyms (‘negative for DVT’, ‘moderate MR’, ‘depressed LVEF’, ‘without r/w/w’). A few acronyms were also marked up for the second run of the system against the test set when they were long and seemed to have no other association in the biomedical literature (‘ARDS’, ‘NTND’) or were extensively used in the training data (‘MR’, ‘TR’, ‘AS’). Entity tagging is case sensitive, thus we marked up certain phrases which were all in capital letters and were commonly observed in the training set (‘ARTERY DISEASE’, ‘PNEUMONIA’). Markup of a few acronyms without a surrounding context, and markup of a few all-caps phrases, constitute the only difference between 1st and 2nd runs of the the system.

(f) Coordination module. This modules marks up noun phrases that are in coordination. This can occur through placement of commas and functional words (‘and’, ‘or’), or through compatible tags in head entities in putative coordinated phrases. Anatomical entities followed by disease tags are united as a single disease/symptom entity. Further anatomical parts in coordinated phrases and followed by a disease tag are also marked up as diseases (‘breast, ovarian and prostate cancer’). Certain disease prefixes are also merged at this stage (‘acute’, ‘lethal’), and bodypart-disease coordination is repeated to detect phrases such as ‘ovarian and early-onset breast cancer’. Certain organism-disease combinations are also detected here (‘viral infection’). We did not make any substantial changes in this module for the CLEF task.

(g) An experimental event detector for the clinical domain. Verbs, gerunds and nominals define ‘trigger words’ which take NP’s as arguments, and define some of the extended annotations in this task. Examples are ‘LV systolic function appears depressed’ and ‘ascending aorta is moderately dilated’. Such extended annotations seemed primarily to correspond to signs and symptoms of disease. We wrote a small module to detect some of these trigger-based sign/symptom events in the testing set. Altogether, about 30 rules were added to detect such events for this task.

3 Results

We first tested the performance of the system as available online [4] against the development sets, and considered only entities marked up as ‘Disease’ or ‘Diseased bodypart’. In the relaxed evaluation mode, precision was 0.91, while recall was 0.51. Accordingly, we modified the system as described in the section above to better capture additional entities in the clinical domain, but without affecting performance in the various other semantic classes detected by the Cocoa tagger.

The major changes that were effected are (a) disambiguation of common words (‘mass’) when they resolve to signs/symptoms in a clinical document and (b) resolution of acronyms that occur commonly in clinical records without an associated expansion. Disambiguation of common words and resolution of acronyms were done in a intrasentential context-sensitive manner based on manual examination of the training set data and appropriate framing of the rules. A small event detection module with about 30 rules was added to detect sentence chunks which corresponded primarily to signs and symptoms (‘hematocrit had not increased’). On the test set, this approach (Run marked ‘TeamRelAgent.1’) yielded a precision of 0.91 and a recall of 0.64, with a f-measure of 0.75.

A number of common short acronyms for diseases/symptoms remained undetected by this approach in the training set. Moreover, there were words and phrases marked all in capital letters in the text that were also left untagged by the system, which is case-sensitive. We added about 70 acronyms that occurred frequently in the training corpus (‘AS’ ‘MVP’, ‘PVD’) as well as 15 all-caps phrases and tried a second run (‘TeamRelAgent.2’) on the test set. The precision lowered by 0.01 to 0.90, but the recall increased far more, from 0.64 to 0.69, with a f-score of 0.78.

4 Discussion

We refined an existing multi-class entity tagger for the biomedical domain (Cocoa) against the test set. The existing tagger already has reasonable performance [4] against a UMLS-based disease corpus (the Arizona disease corpus, [2]). The challenge in extending the system to clinical records was in keeping the precision high while increasing the recall, and yet not compromise performance against other entity classes. With fairly minor improvements, the system achieved a f-score of 0.78 against the test set as compared to the best score of 0.87.

We did not address the problem of increasing precision during this task, apart from addressing obvious errors, such as demarking 'Allergies' when it is a section heading or a department name. Recall was increased by manually analyzing the test for untagged or wrongly tagged entities. Many of these arose from mistagging of common words, such as 'mass' and 'drift', which are symptoms in the clinical context. Acronyms are used frequently in discharge summaries and lab reports without any expansion, and are another source of low recall. Even taking these into account, we could achieve a recall of 0.69 at best in the relaxed evaluation. By comparison, the best-performing system had a recall of 0.83.

Low recall arises for a number of reasons. We did not mark up words such as 'agitated', 'lethargic', 'uncooperative' and 'mass' without an intra-sentential context for disambiguation. We also did not mark up sentence fragments such as 'temperature decreased', as they may not refer to symptoms in other contexts, such as in biochemistry. Rarer acronyms also remain untagged as diseases, as they may refer to chemical or protein names in a general biological context. Even with these constraints, we were able to get a reasonable recall of 0.79 in the training set. However, recall dropped to 0.69 in the test set, for reasons that we have not yet analysed. However, given the small number of modifications that we made to the existing system to increase recall to reasonable figures, we felt that the system is capable of better performance with added effort.

In summary, we have shown that a multi-class entity detection system is capable of achieving reasonable performance in the clinical domain without compromising performance in other classes (data not shown). Clinical documents often contain chemical and protein names and associated quantitative values (e.g. dosage, serum concentrations). A multi-class NER system may thus be useful in correlating multiple entity classes as well as quantitative information with disease occurrence in clinical records. Such correlations would be of relevance in hypothesis-based discovery, such as in cohort analysis, but also in hypothesis-free analysis of large datasets.

Acknowledgments. We thank Dr. K. E. Ravikumar for suggestions and discussions. The ShARe/CLEF eHealth shared task was made possible by NIH grant R01GM090187 to the task organizers.

References

1. Koeling R., Carroll J., Tate A. R., Nicholson A.: Annotating a corpus of clinical text records for learning to recognize symptoms automatically. 2011. Proceedings of Louhi '11.
2. Leaman, R., Miller, C., Gonzalez, G: Enabling Recognition of Diseases in Biomedical Text with Machine Learning: Corpus and Benchmark. 2009. Symposium on Languages in Biology and Medicine, 82-89.
3. Pestian J. P., Brew C., Matykiewicz P., Hovermale D. J., Johnson N., Cohen K. B.: A Shared Task Involving Multi-label Classification of Clinical Free Text. 2007. Association for Computational Linguistics (ACL), 2007:97104.
4. RelAgent Pvt Ltd.: Cocoa. http://npjoint.com/CocoaEval.html
5. Roque, F. S., Jensen P. B., Schmock H., Dalgaard M., Andreatta M., Hansen T., Soeby, K., Bredkjor, S., Juul, A., Werge, T., Jensen L. J., Brunak, S: Using electronic patient records to discover disease correlations and stratify patient cohorts. 2011. PLoS Comp. Bio. 7(8):e1002141.
6. Suominen H., Salantera S., Velupillai S. et al.: Three Shared Tasks on Clinical Natural Language Processing. Proceedings of CLEF 2013. To appear.
7. Uzuner O.: 2011 i2b2/VA co-reference annotation guidelines for the clinical domain. Available from: https://www.i2b2.org/NLP/Coreference/assets/CoreferenceGuidelines.pdf