Cocoa: Extending a rule-based system to tag disease attributes in clinical records

S. V. Ramanan and P. Senthil Nathan

RelAgent Tech Pvt Ltd, Chennai, India,
ramanan,senthil@relagent.com, http://relagent.com

Abstract. We extended Cocoa/Peaberry, our (RelAgent) existing rule based entity and event tagger, to tag attributes associated with diseases in clinical records. The boolean attributes of Negation, Uncertainty and Conditional were handled by an extension of the NegEx algorithm. The multi-valued Course and Severity attributes were detected either within the extended disease spans as output by the system, or by event-based annotation using a predicate-argument framework. The anatomical attribute Body Location was marked up by either finding embedded body parts in the extended disease spans or by being colocated close to the disease span. UMLS IDs for anatomical locations were derived by using a small number of morphological lemmas, and by a few rules derived by manual inspection in case of multiple hits. We used the most frequent value in the training data for Subject, Generic, and time-related attributes.

Keywords: rule-based tagger, disease attributes, clinical notes.

1 Background

Tracking the severity, anatomical location, and temporal factors, including the course, pertaining to a disease/symptom is of significant value in diagnosis. The incidence and progression of a disease are also relevant to tracking response to treatment with various medications, which is useful both in clinical practice as well as in phase trials.

Tagging of diseases, signs and symptoms themselves as well as their normalization to SNOMED terminology was the focus of the 2013 ShARe/CLEF eHealth task 1, which covered a variety of clinical documents, such as discharge summaries and echo/radiology/ECG reports. Disease-tagging tasks share a degree of overlap with previous tasks which marked up radiology reports [5] and discharge summaries [11, 13].

The current ShARe/CLEF eHealth task 2 [2, 3] addresses annotations of the various attributes associated with diseases, signs and symptoms. Specifically, the text span of the disease-related entity as well as its mapping to the SNOMED subset of UMLS are "given" by the organizers, and the performance of systems is evaluated against the detection of only the attributes of these given diseases. The attributes are of several types. Some attributes are boolean; these are the negation, speculation, and conditional attributes. Some attributes are multi-valued; for example, the severity attribute can take "mild", "moderate" and "severe" values. The progression or course attribute takes six values, some pertaining to disease, such as improved/worsened/resolved, which others pertain more to symptoms/signs, such as increased/decreased/changed. The anatomical location of the disease or injury is in an attribute class of its own, with a very large value set, as it is a normalization against the (large) sub-branch of UMLS dealing with body parts. Other annotated attributes are: the bearer of the disease (e.g., the patient or a family member), time-related attributes, and generic symptoms such as fever, which are system-wide and not confined to a discrete anatomic location.

Cocoa/Peaberry is our (RelAgent) existing named entity and event tagger for published literature in the biomedical domain. The system performs reasonably well in various tasks ranging from tagging entities and events in the molecular/cellular domain [8][9] to tagging disease-related entities in the clinical domain [10]. For diseases, the system is designed to tag the maximal compatible span; thus "acute renal insufficiency" would be tagged as a single entity. Thus, many of the attributes required for the current task are already pre-tagged inside the extended entity (location ='renal', severity ='acute') . However, severity/course attributes in some cases ('His condition resolved') are indicated by a verb rather than an adjective, and we use the event-processing capability of the system to tag these cases as well. Further, proximity is used to detect anatomical sites that are distal from the disease mention. Finally, as clinical notes are often syntactically opaque [4], we use a NegEx-based strategy for detecting attributes such as negation and conditionality.

2 System description

We have used the Cocoa/Peaberry system to detect diseases, signs and symptoms for the 2013 CLEF ehealth task 1 [10]. The system is composed of the following modules running in succession: (a) sentence splitter (b) acronym detector (c) POS module (d) word level entity tagger (e) multi word entity detection (e) coordination module (f) shallow parser and (g) predicate-argument detector and finally (h) intra-sentential discourse detection for resolution of argument sharing across predicates. The system detects entities from a range of semantic classes, including proteins, chemicals and clinical procedures in addition to those primary to this task, namely anatomical parts and diseases.

Briefly, the system marks up anatomical parts ('liver') and disease headwords ('cancer') separately, and then merges them to get the extended disease entity. Body parts tagged by the system are composed of maximal spans, such that words corresponding to location ('left upper') or laterality are merged into the body part entity. For disease entities, words describing severity ('acute'), frequency ('recurrent', 'regular'), state ('unresectable', 'disseminated') are also merged if they are located proximally. When anatomical parts are in coordination, and the last occurrence abuts a disease headword (‘liver and breast cancer’), all the coordinated anatomical parts are marked up as disease entities. Finally, unusually named diseases or symptoms such as ‘premature cardiac complex’. ‘long QT syndrome’ are handled. The final tagged disease entity is thus of maximal span in that all adjacent words in any way pertaining to a disease are merged into the entity.

Many disease entities have disjoint spans, especially when they correspond to pathological changes in body parts (‘enlargement of the left ventricle’) or to a source (‘drainage from wound’). In a predicate-argument formalism, one of the disjoint spans appears as an argument of the other span, where the predicate appears either as a verb or in a nominal form. The system has an event detection module to detect such cases.

For the Severity attribute, we used the training set to make a list of trigger words that correspond to the three severity classes, namely ‘slight’, ‘moderate’ and ‘severe’. The disease entity was marked up correspondingly for severity if the extended disease span contained any of these trigger words. Similarity, embedded trigger words associated with the various values for the Course attribute were obtained from the training sets. Examples are ’progressive’ for ’worsened’ and ’healed’ for ’resolved’. However, the majority of the Course attribute data derived from verbal markers for the disease entity. An example for the value ’improved’ is the fragment: ’mental status changes that responded well to Haldol’ which follows the template ’Disease responds to Chemical’. We use the event detection module to mark up the Course attribute for such cases with about 15 trigger word driven event templates (verbs or nominals). However, we ignored the value ’changed’ for the Course attribute as it caused too many false positives, and its occurrence in the training set was small (less than 2%)

As the system marked up anatomical parts before merging them with disease headwords, embedded body locations are automatically output by the system. Additionally the event module also marks up cases where the disease entity is linked to an anatomical part or location through an intervening preposition, as in ‘bleeding in your esophagus’ or ‘loculated effusion seen on the left side’. However, there exist a large number of examples where the anatomical location does not occur in a sentential context, but is implied by the discourse, as when it heads the utterance with a following colon or a hyphen, as in ‘Abdomen: no masses’. Thus for diseases which do not have an embedded or prepositionally proximal body part, we look for occurrences of anatomical locations within 100 characters of the disease entity span. We constructed about 50 rules matching the anatomical part with the disease, e.g., ‘murmur’ or ‘gallop’ match to cardiac entities such as ‘CV’, ‘atheroma’ or ‘extravasation’ match to ‘artery’, ‘clubbing’ and ‘edema’ with ‘extremities’ and so on.

We mapped anatomical entities to UMLS IDs through a collection of morphological transformations which converted the entity as it occurred in the text to a regular expression. The framework for this module is the same as the one we have used in other shared tasks to map diseases to their UMLS or MeSH IDs [7]; both involve substitutions such as modifying ’facial’ to the regex ‘fac(e|ial)’ and ‘ventricular’ to ‘ventric(le|ular)’, apart from generic postpositional changes such as ‘ive’ to ’(ive|ion)’. Altogether, we have 120 rules for such morphological transformations. The regular expression thus constructed was ‘grep’ped against the descriptive phrases for entities in the anatomical subsections of UMLS as given in the task description. Where there were multiple matches, the match with the lowest UMLS ID was chosen, as this was empirically found to best model the training data. An additional set of priority rules were used at the end to reflect certain preferences of the annotators; for example, the UMLS entity ”C0278454|All extremities|extremities” was preferred to ”C0015385|Limb structure|Extremities” when matching the term ”extremities”, while for the term ‘organ’, the UMLS term ’C0229983|Body organ structure’ was preferred to ‘C0178784|Organ’. There were about 130 priority rules for such preferences.

The Negation, Speculation and Conditional attributes were handled by using the NegEx algorithm [1] with additional trigger words as derived from the training data. A few modifications such as the words ‘mild’ and ‘moderate’ limiting the scope of a ’no’ negation to their left were also added. In addition, the event detection module marks negation for any event when the verb/action is conjoined to the appropriate marker (‘not seen’). Beyond these few changes, the NegEx algorithm was used as-is.

We did not address the other attributes. For the Subject and Generic attributes, the data was somewhat sparse, and we decided to leave the default value in place. Detecting time attributes is well known to be difficult task [11], and given our own lack of time, we chose to simply insert the most frequent value in the training set for these attributes, namely ‘none’ for the Temporal Expression attribute and ’overlap’ for the DocTime attribute.

3 Results

We tested and refined the performance of the system against the training set. We then ran two runs against the dataset, which differed only in that the 2nd run split a word into two tokens if there was a slash (‘/’) character in the word. The two runs produced results that did not differ in the overall accuracy (0.843), and we have shown the better of the results of the two runs for the individual attributes in the column titled ‘Test set’ in Table 1 below. The column entitled ‘Best’ is the best result over all systems for each attribute, while the ‘Baseline’ column shows the result if the gold template had been returned unaltered, i.e. the accuracy with the default value for each attribute. These ‘Baseline‘ figures were taken from the accuracy results for systems which had an F-score of 0.0 for that attribute.

One note of relevance is that we did not directly use the gold annotations supplied by the task organizers (except for one case; please see below). Instead, we used the system to itself detect the disease entities while simultaneously marking up the disease attributes. Then, for every disease span in the gold annotations, we found the first disease span in our own annotations that overlapped with the gold span using the overlap algorithm in the 2013 ShARe/CLEF eHealth Task

Table 1. Performance against training and test sets

Attribute	Training set	Test set	Best	Baseline
Overall	0.889	0.843	0.868	–
SV	0.972	0.975	0.982	0.942
CO	0.972	0.963	0.978	0.936
GC	1.000	1.000	1.000	1.000
BL	0.775	0.756	0.797	0.546
DT	0.590	0.024	0.328	–
TE	0.706	0.864	0.864	0.864
NI	0.964	0.944	0.969	–
SC	0.992	0.984	0.995	0.984
UI	0.944	0.955	0.960	0.941
CC	0.976	0.970	0.971	0.961

1 evaluation script [6]. If there was no overlap, we left the attribute template unaltered. If there was an overlap, we copied the attributes from the system-detected disease entity to the gold entity that it overlapped with. However, for the anatomical part attribute alone, we used the algorithms described in the Results section to find any embedded body part or, failing that, the nearest body location compatible with the gold tagged disease entity.

4 Discussion

We extended Cocoa/Peaberry, an existing multi-class entity tagger for the biomedical domain, to detect attributes of disease entities in clinical records. With fairly minor improvements, the system came second in overall accuracy and performed reasonably well in most attributes.

Except for the Body Location attribute, we did not directly use the gold annotated disease entities, instead using the system itself to detect and tag the disease entities and subsequently (and finally) transferring their attributes to overlapping gold annotated entities. We believe therefore that our results on the test set are likely to be close to results against unannotated data, where disease entities are not tagged beforehand, and our results are encouraging in this regard. However, improvement against the baseline is not very high for Cocoa for many attributes, as reflected in the F-scores (not shown), indicating that system performance could benefit from further improvement generally, but particularly for some attributes such as Body Location.

Acknowledgments. We thank Shereen Broido for discussions. The ShARe/CLEF eHealth shared task was made possible by a grant to the task organizers.

References

1. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G., A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. Journal of Biomedical Informatics, 2001. 34: p. 301-310.
2. Elhadad N., Chapman W., O'Gorman T., Palmer M., Savova G.. The ShARe Schema for the Syntactic and Semantic Annotation of Clinical Texts. Under Review.
3. Kelly, L., Goeuriot L., Leroy G., Suominen H., Schreck T., Mowery D.L., Velupillai S., Chapman W.W., Zuccon G., Palotti J. Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. Springer-Verlag.
4. Marsh, E., Sager, N.: Analysis and processing of compact texts. 1982. COLING 82: Proceedings of the Ninth International Conference on Computational Linguistics. North-Holland. 201206.
5. 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.
6. Pradhan, S., Elhadad, N., South, B. R., Martinez, D., Christensen, L., Vogel A., Suominen H., Chapman W. W., Savova, G.. Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. Proceedings of ShARe/CLEF eHealth Evaluation Labs.
7. Ramanan, S. V., Senthil Nathan, P.: Performance of a multi-class biomedical tagger on the BioCreative IV CTD task. 2013. Proceedings of the Fourth BioCreative Challenge Evaluation Workshop vol. 1. Bethesda, MD.
8. Ramanan, S. V., Senthil Nathan, P.: Adapting Cocoa a multi-class entity detector for the CHEMDNER task of BioCreative IV. 2013. Proceedings of the Fourth BioCreative Challenge Evaluation Workshop vol. 2. Bethesda, MD.
9. Ramanan, S. V., Senthil Nathan, P.: Performance and limitations of the linguistically motivated Cocoa/Peaberry system in a broad biomedical domain. 2013. Proceedings of Workshop. BioNLP Shared Task 2013. ACL. Sofia.
10. Ramanan, S. V., Broido S., Senthil Nathan, P. S. V. Ramanan, S. Broido and P. Senthil Nathan: Performance of a multi-class biomedical tagger on clinical records. 2013. Proceedings of ShARe/CLEF eHealth Evaluation Labs.
11. Sun, W., Rushisky, A., Uzuner O.: Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc 2013;0:18.
12. Suominen H., Salantera S., Velupillai S. et al.: Three Shared Tasks on Clinical Natural Language Processing. Proceedings of CLEF 2013. To appear.
13. Uzuner O.: 2011 i2b2/VA co-reference annotation guidelines for the clinical domain. Available from: https://www.i2b2.org/NLP/Coreference/assets/CoreferenceGuidelines.pdf