RelAgent: Entity Detection and Normalization for Diseases in Clinical Records: a Linguistically Driven Approach

Sv Ramanan
RelAgent Tech Pvt Ltd
Adyar, Chennai
India
ramanan@relagent.com

Senthil Nathan
RelAgent Tech Pvt Ltd
Adyar, Chennai
India
senthil@relagent.com

Abstract

We refined the performance of Cocoa/Peaberry, a linguistically motivated system, on extracting disease entities from clinical notes in the training and development sets for Task 7. Entities were identified in noun chunks by use of dictionaries, and events ('The left atrium is dilated') through our own parser and predicate-argument structures. We also developed a module to map the extracted entities to the SNOMED subset of UMLS. The module is based on direct matching against UMLS entries through regular expressions derived from a small set of morphological transformations, along with priority rules when multiple UMLS entries were matched. The performance on training and development sets was 81.0% and 83.3% respectively (Task A), and the UMLS matching scores were respectively 75.3% and 78.2% (Task B). However, the performance against the test set was low by comparison, 72.0% for Task A and 63.9% for Task B, even while the pure UMLS mapping score was reasonably high (relaxed score in Task B = 91.2%). We speculate that our moderate performance on the test set derives primarily from chunking/parsing errors.

1 Introduction

The increasing use of electronic health records, both for satisfying mandatory requirements as well as for administrative reasons, has created a need for systems to automatically tag and normalize disease/sign/symptom mentions. Statistically significant correlations extracted from automated analysis of large databases of clinical records are felt to be useful in detecting phenotype-genotype correlations (reviewed in Kohane (2011)), phenotype-phenotype correlations (Roque et al., 2011) as well as in continuous monitoring of events such as adverse reactions and even early detection of outbreaks of epidemics/infectious diseases (Botsis et al., 2013; Collier, 2012). In this context, Task 7 of SemEval 2014, which is a continuation of the ShARe/CLEF eHealth 2013 task (Pradhan et al., 2013), provides a testbed to evaluate systems that automatically tag and normalize mentions of diseases, signs and symptoms in clinical records, which include discharge summaries and echo, radiology and ECG reports.

Our system consists of (i) Cocoa, a chunk-based entity tagger and (ii) Peaberry, a parser, followed by a module for predicate-argument structure. We have tested the system in a variety of tasks, such as detecting and normalizing mentions of chemicals, proteins/genes, diseases and action terms in the BioCreative 13 Chemdner and CTD tasks (Ramanan and Senthil Nathan, 2013a; Ramanan and Senthil Nathan, 2013b), as well as in detecting cellular and pathological events in the BioNLP cancer genetics task (Ramanan and Senthil Nathan, 2013c); we also participated in the eHealth 2013 task (Ramanan et al., 2013d). Throughout, we have retained a common core platform for simultaneous detection of a multiplicity of entity types as well as for chunking and parsing; we restrict task-specific optimization primarily to post-processing modules. While this strategy may not be optimal

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 477–481,
Dublin, Ireland, August 23-24, 2014.

for any individual task, we feel that it is necessary for multi-document spanning tasks such as literature-based discovery (Swanson, 1988), where connections are established across a variety of scales, e.g. from molecular events to patho-physiological phenotypes. Moreover, these linkages need to be made across a multiplicity of documents from various sources, which encompass a linguistic range from complex syntactical utterances in biomedical publications to free-form phrase-centered clinical notes.

We refined performance against the provided training and development sets, with reasonable performance in Task A (relaxed $f = 0.94$, strict $f = 0.81 - 0.83$, strict recall $0.80 - 0.82$). A module to match text from gold-annotated exact spans to UMLS codes also achieved reasonable performance for Task B (relaxed accuracy $= 0.94-96$). However, the results against from the test set were quite low for Task A, (relaxed $f = 0.87$, strict $f = 0.72$, strict recall $= 0.70$) as well as for Task B (strict $f = 0.64$). Comparatively, the module for UMLS normalization fared better (relaxed $f = 0.91$ in Task B). We speculate that the test set contains entities that are rare in the training/development sets which were chunked incorrectly, and also that the parse errors in the test set arose from syntactic structures missing in the training sets. It is possible that a post-processing statistical module trained on a combination of gold annotations as well as linguistic output may be needed for improving the performance of our system on clinical notes.

2 System description

The basic structure of the entity-tagging system is unchanged from that used in Share/CLEF eHealth 13 (Pradhan et al., 2013) and BioNLP-ST 13. In summary, the system comprises of a sentence splitter, followed by a TBL-based POS tagger and chunker, entity tagging at the single-token level, a module to handle multi-word entities, a noun phrase coordination module, a dependency parser (Ramanan and Senthil Nathan, 2013c), and finally a semantic module to tag disease-related events.

The generic system has dictionaries and morphological rules for detecting diseases and body parts. However, there are many extensions needed for clinical notes, which (i) make extensive use of common words and phrases for describing symptoms, which requires word sense disambiguation, (ii) use unusual phrases for signs and symptoms and (iii) are full of undefined acronyms. We isolated such specialization to disease-related entities within noun phrases in clinical documents inside a subroutine in the multi-word tagger module. These were identified by a frequency-based analysis of words and phrases in the training and development corpora. Thus, a few ambiguous words and phrases such as 'crackles', 'complaints', 'mass effect' and 'focal consolidation' were tagged as disease markers regardless of context. Generally, however, even common clinical words such as 'redness' and 'swelling' were tagged only in the presence of neighboring context words. The appearance of major body parts such as 'Abdomen', 'Neck, 'Extremities' at the beginning of a line followed by a colon or a hyphen was taken as a discourse reference marker for the rest of the line to tag acronyms such as 'NT/ND' and dangling adjectives such as 'soft' and 'warm'. Very common acronyms (≈ 100) both for anatomical parts ('LUQ') and diseases ('DMII') were also tagged inside the specialized subroutine, as were common abbreviations ('regurg' for regurgitation) and words with common spelling errors. Finally, some event/process words which we found to almost always represent clinical conditions in the training text were tagged as disease markers. Examples are 'aspiration', agitation' and 'confusion'.

We also extended our generic event processing module with a task-specific routine to take into account descriptions of (mostly) signs/symptoms specific to clinical documents. These fall into several categories: (i) abnormal changes in body parts or organ systems, such as 'The left atrium was moderately enlarged', 'Nose is bloody' and 'redistribution of pulmonary blood flow' (ii) symptoms such as 'The patient was unable to walk', 'His speech was slurred', 'He had difficulty breathing' and 'alteration of consciousness' (iii) changes in parameters marked by phrases/clauses such as 'elevation of troponin', 'QR interval was prolonged' and 'decreased blood sugar'. Certain environmental conditions such as 'exposure to asbestos' were also handled. Finally, events with a default animate theme were tagged regardless of their actual arguments to handle sentences/phrases where our syntax module failed to extract the correct theme or the theme is to be inferred from the discourse; the ≈ 40 words in this set included verbs such as 'vomit', 'shivering', 'lethargic', 'violent' and 'somnolent'.

The above treatment served to demarcate spans for diseases that overlap with the gold annotations. The system merges words/phrases denoting a body part with adjoining words that denote diseases, and also merges words denoting severity into the disease span, since our system design strategy was to generate the longest contiguous span that can refer to a disease. However, the primary score in the shared task are with respect to exact matches with the gold annotations. We therefore wrote a small post-processing module to omit words in an approximate match that refer to severity ('acute') as well as to excise phrases dealing with intra-organ parts or their location (such as 'lobes' or 'left/right') - such words/phrases are usually omitted from the UMLS descriptions of diseases to which the gold annotations hew closely. Also, we noticed that certain words such as 'wounds' and 'lesions' do not embed an anatomical entity within their description in the gold annotations. Yet another point is that, while parameters are marked up as indicative of a symptom only when they take on abnormal values ('elevated LDL'), the direction of change is almost always omitted from the gold annotations. Descriptors of the patient ('He') are also excised. Altogether, we constructed about 40 rules to trim the approximate span into one more conformant to the exact form in the gold annotations.

Task B requires mapping diseases phrases into the SNOMED subset of UMLS as specified in the task description. We proceeded on the assumption that the exact (gold) entity spans were constructed by annotators to closely map into the UMLS descriptions. Accordingly, we used the text as defined by the gold spans and attempted to map them directly into the UMLS definitions after some preprocessing steps that constructed a regular expression: (a) common spelling errors were corrected (b) body part and disease acronyms were expanded (c) common variants were added as alternates i.e. 'tumou?rs?' were expanded into '(tumou?r|neoplasm|carcinoma)s?' (d) adjectival and nominal variants were added e.g. both 'atrium' and 'atrial' were converted into '(atri)(al?|um)', and more generally, adjectival endings were generalized, for example, the ending 'ic' was converted into '(i[ac]|ism)'. (e) singular and plural forms were converted into choices e.g. 'artery' was rendered as 'arter(y|ies)'.

Altogether, we have ≈ 120 rules for variant morphological forms, covering adjectives, nouns and number. The resulting regular expression was directly matched (using 'grep') against UMLS text entries. Generally, several matches were found. Matches against the defining entry (the first one) were prioritized, otherwise the entry with the largest CUID was taken. Finally, we noted that some UMLS CUID's were preferred to others; for example, 'C0007115 - Malignant neoplasm of thyroid' is preferred to 'C0549473 - Thyroid carcinoma'. The preferred choices were inferred from gold annotation frequencies, and correspond to ≈ 100 remapping rules.

3 Results and Discussion

With a few minor changes to the system used in the Share/CLEF 2013, we obtained a relaxed f-measure in Task A of 0.88 in the training and development sets. Thereafter we alternately refined performance in Task A against the provided training set using the development set as a testbed, or vice versa. As described in the last section, these refinements took the form of adding context-sensitive rules for disease-related words and phrases in order of their frequencies in the training/development sets. While we could thereby improve performance against both training and development sets (relaxed $f = 0.94$), we noticed that improvements in the performance against the training set did not correlate with better performance against the development set and vice versa, probably implying that 6% or more of the entities are unique to each set, or that we were unable to catch similarities. A similar orthogonal situation resulted in our attempt to improve performance against exact matches on the training and development sets, strict $f = 0.81 - 0.83$, strict recall $0.80 - 0.82$. The observation of orthogonal entity sets in different datasets for about 6% of entities is seemingly validated in the test set, where the results showed a relaxed $f = 0.87$, which is quite close to the baseline performance (0.88 in the Share/CLEF 2013 task); the highest scoring system had relaxed $f = 0.91$ by comparison. We speculate that our insistence on contextual clues for entity tagging is another cause for low relaxed performance on the test set.

Performance of the system for exact matches on the test set (strict $f = 0.72$) suffered greatly in comparison to the training/development sets. This could be partly ascribed to the 7% lower performance on the relaxed f-score (i.e. we missed many entities altogether) from 0.94 in training/development sets to 0.87 in the test set. Even accounting for this, there is an additional performance drop of about $3-4\%$ in exact match on the test set compared to training/development sets. One implication is that that our rule-base method for pruning approximate matches to exact spans is probably sub-optimal, and should be supplemented or replaced by a statistical algorithm. As noted earlier, gold annotations are probably made by annotators with respect to UMLS definitions, and have some degree of arbitrariness associated with them depending on the granularity of the UMLS definition e.g. in the choice of whether to remove or retain a body location in the gold span. Given the size of the UMLS definition set, a statistical approach is probably likely to do better than a rule-based system in the task of reducing approximate matches to exact spans.

The poor performance in Task A (strict recall = 0.70) directly impinges on our low 'strict' score in Task B (= 0.64); this score is simply a product of the strict recall in Task A and the accuracy of mapping to UMLS, where the latter score is given by the Task B 'relaxed' score (= 0.91). An interesting feature is the mapping accuracy for our system on the test set suffered a relatively small drop when compared to the mapping accuracies on the training and development sets, which were 0.94 and 0.96 respectively. We interpret this reasonably high figure for the mapping score (the best among the top 10+ teams in Task B) as validation of our hypothesis that gold annotations are made with respect to UMLS definitions, which also strengthens the case (made above) for the need to incorporate a (semi-)statistical approach for pruning overlap matches to exact matches in our system.

Clinical documents are terse and full of phrasal observations and incomplete sentences, often with missing punctuation. We have adapted a linguistically based system to detect disease-related entities and events with moderate performance; our observation on the training/development sets is that most errors arise from parsing/ chunking errors on grammatically incomplete phrases. The second task, namely mapping disease-related entities/events to SNOMED/UMLS, requires tagged entity spans to correspond closely to UMLS definitions; system performance in this regard can probably be usefully supplemented by statistical approaches. Given proper entity spans, a small set of morphological transformations gives high performance in mapping to UMLS ID's. We speculate that a chunk-annotated corpus of clinical records may help in improving performance for linguistically derived systems.

References

Isaac S. Kohane. 2011. *Using electronic health records to drive discovery in disease genomics.* Nat Rev Genet. 2011 Jun;12(6):417-28.

Francisco S. Roque, Peter B. Jensen, Henriette Schmock, Marlene Dalgaard, Massimo Andreatta, Thomas Hansen, Karen Soeby, Soren Bredkjor, Anders Juul, Thomas Werge, Lars J. Jensen and Soren Brunak. 2011. *Using electronic patient records to discover disease correlations and stratify patient cohorts.* PLoS Comp. Bio. 7(8):e1002141.

Sv Ramanan and Senthil Nathan. 2013. *Performance of a multi-class biomedical tagger on the BioCreative IV CTD task.* Proceedings of the Fourth BioCreative Challenge Evaluation Workshop vol. 1. Bethesda, MD.

Sv Ramanan and Senthil Nathan. 2013. *Adapting Cocoa a multi-class entity detector for the CHEMDNER task of BioCreative IV.* Proceedings of the Fourth BioCreative Challenge Evaluation Workshop vol. 2. Bethesda, MD.

Sv Ramanan and Senthil Nathan. 2013. *Performance and limitations of the linguistically motivated Cocoa/Peaberry system in a broad biomedical domain.* Proceedings of Workshop. BioNLP Shared Task 2013. ACL. Sofia.

Sv Ramanan, Shereen Broido and Senthil Nathan. 2013. *Performance of a multi-class biomedical tagger on clinical records.* Proceedings of ShARe/CLEF eHealth Evaluation Labs.

Don R. Swanson. 1988. *Migraine and Magnesium: Eleven Neglected Connections.* Persp. Bio. Med. 31(4), 526-557.

Sameer Pradhan, Noemie Elhadad, Brett R. South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W. Chapman and Guergana Savova. 2013. *Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop.* Proceedings of ShARe/CLEF eHealth Evaluation Labs, 23-26 September, Valencia, Spain

Taxiarchis Botsis , Michael D. Nguyen , Emily J. Woo, Marianthi Markatou and Robert Ball. 2011. *Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection.* J Am Med Inform Assoc. 2011 Sep-Oct;18(5):631-8

Nigel Collier. 2012. *Uncovering text mining: A survey of current work on web-based epidemic intelligence.* Glob Public Health. Aug 2012; 7(7): 731-749.