CUAB: Supervised Learning of Disorders and their Attributes Using Relations

James Gung
University of Colorado
1111 Engineering Drive
Boulder, Colorado 80309, USA
james.gung@colorado.edu

John David Osborne, Steven Bethard
University of Alabama at Birmingham
1720 2nd Ave South
Birmingham, AL, USA 35294
{ozborn,bethard}@cis.uab.edu

Abstract

We implemented an end-to-end system for disorder identification and slot filling. For identifying spans for both disorders and their attributes, we used a linear chain conditional random field (CRF) approach coupled with cTAKES for pre-processing. For combining disjoint disorder spans, finding relations between attributes and disorders, and attribute normalization, we used l2-regularized l2-loss linear support vector machine (SVM) classification. Disorder CUIs were identified using a back-off approach to YTEX lookup (CUAB1) or NLM UTS API (CUAB2) if the target text was not found in the training data. Our best system utilized UMLS semantic type features for disorder/attribute span identification and the NLM UTS API for normalization. It was ranked 12th in Task 1 (disorder identification) and 6th in Task 2b (disorder identification and slot filling) with a weighted F Measure of 0.711.

1 Introduction

One of the core problems in the field of clinical text processing is the identification and normalization of medical disorders (Pradhan et al., 2014). A secondary problem is the identification of attributes for the identified disorders such as their severity or body location. Attribute identification and normalization helps to better describe the disorder context, allowing for a better determination of the appropriateness of the discovered disorder for the task at hand.

SemEval-2015 Task 14 addresses these problems as separate tasks, assessing end to end systems capable of identifying both disorders and attributes from unlabeled clinical text. The first task requires participants to identify discontinuous disorder spans in clinical text and normalize them to a UMLS Concept Unique Identifier (CUI) that is both within the disorder Semantic Group and present in SNOMED CT. The second task requires identification of disorder CUIs as well as 8 additional attributes associated with each disorder as shown in Table 1 on the shared task page[1]. For each attribute, the span offset of the lexical cue must also be identified, which may be discontinuous.

2 Approach

We combined and extended our previous work (Gung, 2014; Osborne et al., 2014) for the ShARe/CLEF 2013 eHealth Evaluation Lab (Suominen et al., 2013). Both previous systems and our base system for this task are based on the clinical Text Analysis and Knowledge Extraction System (cTAKES) (Savova et al., 2010), an open source pipeline for the natural language processing (NLP) of clinical text that utilizes the Unstructured Information Management Architecture (UIMA) (Ferrucci and Lally, 2004) framework. Our combined system is available for download at https://github.com/jgung/ClearClinical.

We developed two systems for this task that differed in their method of CUI lookup and the presence of UMLS semantic type features. The first system (CUAB1) uses YTEX (Garla et al., 2011) to disambiguate CUIs returned from the cTAKES dictionary

[1] http://alt.qcri.org/semeval2015/task14/index.php?id=task-description

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 417–421,
Denver, Colorado, June 4-5, 2015.

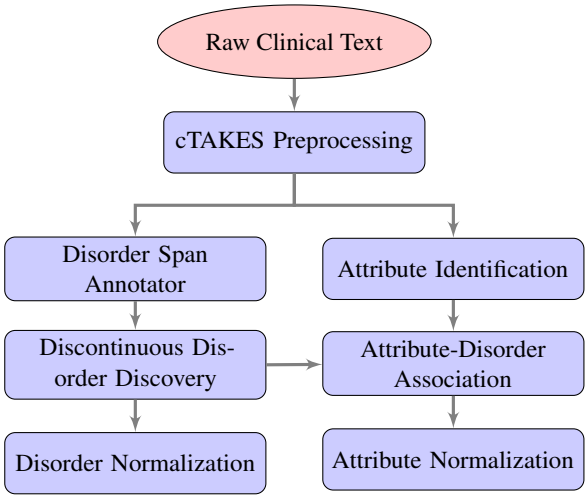

Figure 1: Pipeline of the system components

annotator. The second system (CUAB2) uses UMLS Terminology Services (UTS) for the same task and additional UMLS features for disorder/attribute span annotation. For both of these systems, we relied on cTAKES for pre-processing, using the default pipeline from the cTAKES ClinicalPipelineFactory class to perform tokenization, sentence segmentation, part of speech (POS) tagging and chunking.

2.1 Task 1 - Disorder Identification and Normalization

We broke down Task 1 into 3 different tasks as shown in Figure 1: identification of disorder spans, linking of disjoint disorder spans into single discontinuous disorders, and association of the final (dis)continuous disorder spans with CUIs.

2.1.1 Disorder Span Annotation

Span identification in Task 1 was accomplished with the same begin-inside-outside (BIO) token classification methodology as in previous work (Gung, 2014) but using the updated training data. Spans of putative disorders were labeled using a linear chain CRF with features identical to those used in previous work. Examples of these features are shown in Table 1. The disorder span tagger was implemented using the ClearTK machine learning framework (Bethard et al., 2014) which presents a UIMA interface for machine learning models and wraps classifiers such as CRFSuite (Okazaki, 2007).

Feature Type	Example Feature
Token	First token of each of the two annotations
POS	Part-of-speech tags (e.g, NN) of each of the two annotations
Phrase-chunk	Phrase chunks (e.g., NP, VP) between the two annotations
Dependency path	Max distance to common ancestor of the two annotations
Dependency tree	Concatenation of head word and governing word for each of the two annotations
Named entity	Number of named entity mentions between the two annotations

Table 1: Feature types and examples for features used to associated disjoint spans into a discontinuous disorder and to associate attributes with a candidate disorder

2.1.2 Discontinuous Disorder Discovery

In a departure from previous work (Gung, 2014), we trained our own relation extractor for the discovery of discontinuous spans, rather than relying on existing models used by ClearNLP's SRL system and the cTAKES relation extractor. We used a l2-regularized l2-loss linear SVM classifier (via the ClearTK wrapper to LibLinear) to predict when two disorder spans identified in the previous step should be combined into a single disorder. We used a subset of features from the cTAKES relation extractor including token features (e.g., last word in disjoint span), POS features, phrase chunks (e.g., phrase chunk between first head), dependency tree information (e.g., dependencies on POS tags, words), dependency path information (e.g., mean distance to common ancestor) and the number of named entities between the disjoint spans. A list of these features with examples is shown in Table 1 and more interested readers can review the source code made available.

We explored some additional features to improve span detection including pointwise mutual information from the provided unlabeled MIMIC notes and CUI-normalized segment header information. Neither feature provided a performance improvement on the training data and thus they were excluded from our final systems.

System	Rank	TP	FP	FN	P	R	F
Strict Results							
CUAB1	23	3514	1381	2634	0.718	0.572	0.636
CUAB2	12	4202	1516	1946	0.735	0.683	0.708
ezDI	1	-	-	-	0.783	0.732	0.757
Relaxed Results							
CUAB1	-	3632	1263	2516	0.742	0.591	0.658
CUAB2	-	4357	1361	1791	0.762	0.709	0.734
ezDI	-	-	-	-	0.815	0.761	0.787

Table 2: Performance on disorder identification and normalization (Task 1), including rank among the 39 competing systems (Rank), true positives (TP), false positives (FP), false negatives (FN), precision (P), recall (R) and F-measure (F). Task ranking was only given for strict scoring.

2.1.3 Disorder Normalization

Disorder normalization in both systems used a dictionary of text-to-CUI mappings from the training data as the primary attempt to normalize the disorders. In CUAB2, any text not normalized by this training dictionary was assigned a CUI using UMLS UTS web services whereas in CUAB1 the assignment was made using the cTAKES dictionary annotator with YTEX to resolve ambiguous terms. In both systems text that failed all of these methods was designated as CUI-less.

2.2 Task 2 - Attribute Identification and Normalization

We broke this task down into 3 different steps as shown in Figure 1: detection of attribute spans, association of those spans to the disorders already identified, and the normalization of the attribute spans (slot filling).

2.2.1 Attribute Identification

To detect attribute spans we used the same linear chain CRF model with the same features that we used to detect disorder spans in Task 1.

As in disorder identification, we labeled tokens as either the beginning, inside, or outside (BIO) of an attribute. Contiguous non-outside chunks were assembled and marked as possible candidate attributes.

2.2.2 Associating Attributes with Disorders

We again used a l2-regularized l2-loss linear SVM classifier model to link our candidate attributes to the disorders discovered by our system in Task 1. This classifier used the same feature set as was used for merging disorder spans (see Table 1).

System	Accuracy
YTEX	0.650
UTS	0.644

Table 3: Accuracy of Disorder Normalization on Training Data

2.2.3 Attribute Normalization

Attributes for disorders were normalized using a l2-regularized l2-loss linear SVM classifier using as features the full text of the attribute, the text of the tokens within the attribute annotation, and the text of the tokens appended with the attribute type.

3 Results

3.1 Task 1

Table 2 shows the performance of the CUAB systems on disorder identification and normalization (Task 1), as well as the performance of the top system in the shared task. The best CUAB system (CUAB2) used UMLS semantic type features for disorder span identification and UMLS Terminology Services (UTS) for CUI lookup and ranked 12th out of 39 systems, achieving precision and recall that were both about 0.05 below the top system. CUAB1 was ranked 23rd but not because the system was less able to normalize disorder CUIs. As shown by the training data in Table 3, both UTS and YTEX had similar accuracy in predicting CUIs. A more plausible explanation for the relatively higher performance of CUAB2 is a result of more accurate span detection due to its

System	Rank	TP	FP	FN	P	R	F	A	WA	F*A	F*WA
CUAB1	17	4627	258	1521	0.947	0.753	0.839	0.873	0.669	0.732	0.561
CUAB2	6	5376	328	772	0.942	0.874	0.907	0.908	0.784	0.824	0.711
UTH-CCB	1	-	-	-	-	-	0.926	0.941	0.873	0.871	0.808

Table 4: Performance on disorder identification, normalization and slot-filling (Task 2b), including rank among the 23 competing systems (Rank), true positives (TP), false positives (FP), false negatives (FN), precision (P), recall (R), F-measure (F), accuracy (A), weighted accuracy (WA).

Slot	CUAB1	CUAB2
BodyLoc	-	0.656
Disorder CUI	0.783	0.808
Conditional	-	0.661
Course	-	0.773
Generic	-	0.885
Negation	-	0.850
Severity	-	0.861
Subject	-	0.846
Uncertainty	-	0.750

Table 5: Weighted accuracy by attribute type on slot-filling

incorporation of additional UMLS lookup features for span detection that were unintentionally left absent in CUAB1. Given the nearly identical results in training between UTS and YTEX, the much better performance of CUAB2 in Task 1 is best explained by the importance of vocabulary features in disorder normalization. Unfortunately the test dataset is not available for us to re-run and confirm this.

Table 4 shows the performance of the CUAB systems on the combined task of disorder identification, normalization and slot-filling (Task 2b). The best CUAB system (CUAB2) again used UMLS features for disorder span and attribute annotation and UTS for CUI lookup and ranked 6th out of 23 systems, achieving an F-measure, accuracy and weighted accuracy about 0.02, 0.03 and 0.09, respectively, below the top system.

Table 5 shows the performance of the CUAB systems broken down by attribute type. The CUAB1 system made only disorder predictions for Task 2b, hence all other results are omitted.

4 Discussion

One strength of our system is that it took exactly the same approach (classifier and feature set) to the problem of merging disjoint disorder spans and the problem of associating attributes with disorder mentions. Our CUAB2 system ranked well and was close to the top systems, which suggests that treating these two problems in the same way was a reasonable approach. This lends credence to the notion that deriving new features for either the merging of disjoint disorder spans or the association of attributes with disorders could be useful for either problem.

One issue of concern is that the accuracy of CUI prediction is still very dependent on training data. Our submitted systems used a direct string lookup from a dictionary built on the training data, before falling back to UTS or YTEX if the example was not found in the training data. This approach achieved a disorder CUI accuracy of up to nearly 81%. But when the training data isn't used for CUI identification, as shown in an experiment on the task training data (Table 3), we only achieve about 65% accuracy. This suggests that approximately 15%+ additional accuracy is entirely a result of having already seen the concept in the training data and that our system (and others relying on the training data) would likely see close to a 15% drop off in disorder CUI prediction accuracy when applied to a new medical sub-domain.

Our scheme uses two classifiers, one to detect and another to merge entities. Future work may include investigating the possibility of employing a single classifier with a more complex tagging schema than BIO to perform these tasks jointly.

Acknowledgments

This work was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number UL1TR00165. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Steven Bethard, Philip Ogren, and Lee Becker. 2014. Cleartk 2.0: Design patterns for machine learning in uima. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3289–3293, Reykjavik, Iceland, 5. European Language Resources Association (ELRA). (Acceptance rate 61%).

David Ferrucci and Adam Lally. 2004. Uima: an architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, 10(3-4):327–348.

V. Garla, V.L. Re III, Z. Dorey-Stein, F. Kidwai, M. Scotch, J. Womack, A. Justice, and C. Brandt. 2011. The yale ctakes extensions for document classification: architecture and application. *Journal of the American Medical Informatics Association*, 18(5):614–620.

James Gung. 2014. Using relations for identification and normalization of disorders: Team clear in the share/clef 2013 ehealth evaluation lab.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs). *URL http://www. chokkan. org/software/crfsuite*.

John David Osborne, Binod Gyawali, and Thamar Solorio. 2014. Evaluation of ytex and metamap for clinical concept recognition. *arXiv preprint arXiv:1402.1668*.

Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014. Semeval-2014 task 7: Analysis of clinical text. *SemEval 2014*, 199(99):54.

G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, and C.G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

H Suominen, S Salantera, S Velupillai, et al. 2013. Three shared tasks on clinical natural language processing. In *Proceedings of the Conference and Labs of the Evaluation Forum*.