UWM: Disorder Mention Extraction from Clinical Text Using CRFs and Normalization Using Learned Edit Distance Patterns

Omid Ghiasvand
University of Wisconsin-Milwaukee
Milwaukee, WI
ghiasva2@uwm.edu

Rohit J. Kate
University of Wisconsin-Milwaukee
Milwaukee, WI
katerj@uwm.edu

Abstract

This paper describes Team UWM's system for the Task 7 of SemEval 2014 that does disorder mention extraction and normalization from clinical text. For the disorder mention extraction (Task A), the system was trained using Conditional Random Fields with features based on words, their POS tags and semantic types, as well as features based on MetaMap matches. For the disorder mention normalization (Task B), variations of disorder mentions were considered whenever exact matches were not found in the training data or in the UMLS. Suitable types of variations for disorder mentions were automatically learned using a new method based on edit distance patterns. Among nineteen participating teams, UWM ranked third in Task A with 0.755 strict F-measure and second in Task B with 0.66 strict accuracy.

1 Introduction

Entity mention extraction is an important task in processing natural language clinical text. Disorders, medications, anatomical sites, clinical procedures etc. are among the entity types that predominantly occur in clinical text. Out of these, the Task 7 of SemEval 2014 concentrated on extracting (Task A) and normalizing (Task B) disorder mentions. Disorder mention extraction is particularly challenging because disorders are frequently found as discontinuous phrases in clinical sentences. The extracted mentions were then to be normalized by mapping them to their UMLS CUIs if they were in the SNOMED-CT part of UMLS and belonged to the "disorder" UMLS semantic group, otherwise they were to be declared as "CUI-less". This normalization task is challenging because disorder names are frequently mentioned in modified forms which prevents their exact matching with concept descriptions in UMLS.

Our team, UWM, participated in both Task A and Task B. We modelled disorder mention extraction as a standard sequence labeling task. The model was trained using Conditional Random Fields (Lafferty et al., 2001) with various types of lexical and semantic features that included MetaMap (Aronson, 2001) matches. The model was also inherently capable of extracting discontinuous disorder mentions. To normalize disorder mentions, our system first looked for exact matches with disorder mentions in the training data and in the UMLS. If no exact match was found, then suitable variations of the disorder mentions were generated based on the commonly used variations of disorder mentions learned from the training data as well as from the UMLS synonyms. We developed a novel method to automatically learn such variations based on edit distances (Levenshtein, 1966) which is described in the next section.

Our Team ranked third on Task A and second on Task B in the official SemEval 2014 Task 7 evaluation (considering only the best run for each team). We also present results of ablation studies we did on the development data in order to determine the contributions of various features and components of our system.

2 Methods

2.1 Task A: Disorder Mention Extraction

We modelled disorder mention extraction as a sequence labeling task with the standard "BIO" (Begin, Inside, Outside) scheme of output labels for sentence tokens. The tokens labelled "I" following the latest "B" token are extracted together as

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 828–832,
Dublin, Ireland, August 23-24, 2014.

a disorder. For example, in the following labelled sequence "the/O left/B atrium/I is/O moderately/O dilated/I", "left atrium dilated" will be extracted as a disorder. The labeling scheme thus naturally models discontinuously mentioned disorders which is one challenging aspect of the disorder mention extraction task.

The sequence labeling model is trained using Condition Random Fields (CRFs) (Lafferty et al., 2001) using the five group of features shown in Table 1. The clinical reports are first pre-processed using Stanford CoreNLP[1] for tokenization, sentence segmentation and part-of-speech (POS) tagging which help in obtaining the lexical features (Group 1). The semantic features (Group 2) are obtained by matching the tokens, along with bigrams and trigrams in UMLS. For the first three features in Group 2, only the eleven semantic types under the "disorder" semantic group are considered.[2] If a token is a concept in UMLS with "disorder" semantic group then its feature is assigned the value of its semantic type (for example "congenital abnormality", "Neoplastic process", etc.) otherwise it is assigned the value "Null". The next three features in Group 2 take Boolean values depending upon whether the bigram or trigram is present in UMLS as a concept or not. The last feature in Group 2 takes CUI as its value if the word is a concept in UMLS otherwise it takes "Null" as the value.

The features in Group 3 are obtained by running MetaMap (Aronson, 2001). The lemmatized version of word obtained using Stanford CoreNLP is used as an additional feature in Group 4. Finally, if the word is an abbreviation according to a list of clinical abbreviations[3] then its full-form is obtained.[4] The full-form, whether it is in UMLS, and its semantic type (out of "disorder group") are used as features under Group 5. We used the CRFsuite (Okazaki, 2007) implementation of CRFs.

Group 1: Lexical
Word
Next word
Previous word
POS tag of word
POS tag of next word
POS tag of previous word
Next to next word
Previous to previous word
Length of the word
Group 2: Semantic
UMLS semantic type of word
UMLS semantic type of next word
UMLS semantic type of previous word
Bigram with next word is in UMLS
Reverse bigram with next word is in UMLS
Trigram with next two words is in UMLS
CUI of the word
Group 3: MetaMap
Word tagged as disorder by MetaMap
Next word tagged as disorder by MetaMap
Previous word tagged as disorder by MetaMap
Group 4: Lemmatization
Lemmatized version of the word
Group 5: Abbreviation
Full-form
Full-form is in UMLS
UMLS semantic type of full-form

Table 1: Features used to train the CRF model for disorder mention extraction.

2.2 Task B: Disorder Mention Normalization

The extracted disease mentions from Task A are normalized in Task B as follows. As a first step, our system tries to exactly match the disease mentions in the training data. If they match, then the corresponding CUI or CUI-less is the output. If no match is found in the training data, then the system tries to exactly match names of concepts in UMLS including their listed synonyms.[5] If a match is found then the corresponding CUI is the output. If the mention does not match either in the training data or in the UMLS and if it is an abbreviation according to the abbreviation list (same as used in Task A), then its full-form is used to exactly match in the training data and in UMLS. However, what makes the normalization task challenging is that exact matching frequently fails. We employed a novel method that learns to do approximate matching for this task.

We found that most failures in exact matching were because of minor typographical variations due to morphology, alternate spellings or typos. In order to automatically learn such variations, we developed a new method based on edit distance which is a measure of typographical similarity between two terms. We used a particular type of well-known edit distance called Levenshtein dis-

[1] http://nlp.stanford.edu/software/corenlp.shtml

[2] We found that using all semantic groups negatively affected the performance.

[3] http://en.wikipedia.org/wiki/List_of_medical_abbreviations

[4] If multiple full-forms were present then only the first one was used. In the future, one could improve this through abbreviation disambiguation (Xu et al., 2012).

[5] In accordance to the task definition, only the concepts listed in SNOMED-CT and of the UMLS semantic group "disorder" are considered in this step.

Learned Edit Distance Pattern	Comments
SAME o INSERT u SAME r	Change American spelling to British
INSERT s SAME space	Pluralize by adding "s" before space
DELETE i DELETE e SUBSTITUTE s/y	Example: "Z-plasties" → "Z-plasty"
START SAME h INSERT a SAME e SAME m SAME o	Variation: "hemo..." → "haemo..."
DELETE space DELETE n DELETE o DELETE s END	Drop " nos" in the end
SAME s SUBSTITUTE i/e SAME s	Example: "metastasis" → "metastases"

Table 2: A few illustrative edit distance patterns that were automatically learned from UMLS and the training data.

Data used for training	Task A						Task B	
	Strict			Relaxed			Strict	Relaxed
	P	R	F	P	R	F	Accuracy	Accuracy
Training + Development	0.787	0.726	0.755	0.911	0.856	0.883	0.660	0.909
Training	0.775	0.679	0.724	0.909	0.812	0.858	0.617	0.908

Table 3: SemEval 2014 Task 7 evaluation results for our system. Precision (P), recall (R) and F-measure (F) were measured for Task A while accuracy was measured for Task B.

tance (Levenshtein, 1966) which is defined as the minimum number of edits needed to convert one term into another. The edits are in the form of insertions, deletions and substitution of characters. For example, the term "cyanotic" can be converted into "cyanosis" in minimum two steps by substituting "t" for "s" and "c" for "s", hence the Levenshtein edit distance between these terms is two. There is a fast dynamic programming based algorithm to compute this. The algorithm also gives the steps to change one term into another, which for the above example will be "START SAME c SAME y SAME a SAME n SAME o SUBSTITUTE t/s SAME i SUBSTITUTE c/s END". We will call such a sequence of steps as an *edit distance pattern*.

Our method first computes edit distance patterns between all synonyms of the disorder concepts is UMLS[6] as well as between their mentions in the training data and the corresponding tagged concepts in UMLS. But these patterns are very specific to the terms they are derived from and will not directly apply to other terms. Hence these patterns are generalized next. We define generalization of two edit distance patterns as their largest contiguous common part that includes all the edit operations of insertions, deletions and substitutions (i.e. generalization can only remove "SAME", "START" and "END" steps). For example, the generalized edit distance pattern of "cyanotic → cyanosis" and "thrombotic → thrombosis" will be "SAME o SUBSTITUTE t/s SAME i SUBSTITUTE c/s END", essentially meaning that a term that ends with "otic" can be changed to end with "osis". Our method generalizes every pair of edit distance patterns as well as repeatedly further generalizes every pair of generalization patterns.

Not all generalization patterns may be good because some may change the meaning of terms when applied. Hence our method also evaluates the goodness of these patterns by counting the number of *positives* and *negatives*. When a pattern is applied to a UMLS term and the resultant term has the same CUI then it is counted as a positive. But if the resultant term has a different CUI then it is counted as a negative. Our system heuristically only retains patterns that have the number of positives more than the number of negatives and have at least five positives. Our method learned total 554 edit distance patterns, Table 2 shows a few illustrative ones.

These patterns are used as follows to normalize disease mentions. When exact matching for a disease mention in the training data and the UMLS fails, then our system generates its variations by applying the learned edit distance patterns. These variations are then searched for exact matching in the UMLS. If even the variations fail to match then the variations of possible full-forms (according to the abbreviation list) are tried, otherwise the mention is declared CUI-less. Note that while our method learns variations only for disorder mentions, it is general and could be used to learn variations for terms of other types. Finally, because it is a learning method and it also learns variations used in the training data, it is capable of learning variations that are specific to the style or genre of the clinical notes that constitute the training data. We note that the problem of matching variations is analogous to the duplicate detection problem

[6]Due to the large size of UMLS, we restricted to the second of the two concept files in the 2013 UMLS distribution.

in database records (Bilenko and Mooney, 2003). But to the our best knowledge, no one has used an approach to learn patterns of variations based on edit distances. We used the edit-distance patterns only for Task B in this work, in future we plan to also use them in Task A for the features that involve matching with UMLS.

3 Results

The organizers of the SemEval 2014 Task 7 provided the training, the development and the test data containing 199, 99 and 133 clinical notes respectively that included de-identified discharge summaries, electrocardiogram, echocardiogram and radiology reports (Pradhan et al., 2013). The extraction performance in Task A was evaluated in terms of precision, recall and F-measure for strict (exact boundaries) and relaxed (overlapping boundaries) settings. The normalization performance in Task B was evaluated in terms of strict accuracy (fraction of correct normalizations out of all gold-standard disease mentions) and relaxed accuracy (fraction of correct normalizations out of the correct disease mentions extracted in Task A). Note that a system's strict accuracy in Task B depends on its performance in Task A because if it misses to extract a disease mention in Task A then it will get zero score for its normalization.

Table 3 shows the performance of our system as determined through the official evaluation by the organizers. The systems were evaluated on the test data when trained using both the training and the development data as well as when trained using just the training data. When trained using both the training and the development data, our team ranked third in Task A and second in Task B considering the best run of each team if they submitted multiple runs. The ranking was according to the strict F-measure for Task A and according to the strict accuracy for Task B. When trained using just the training data, our team ranked second in Task A and first in Task B.

We also performed ablation study to determine the contribution of different components of our system towards its performance. Since the gold-standard annotations for the test data were not made available to the participants, we used the development data for testing for the ablation study. Table 4 shows the results (strict) for Task A when various groups of features (shown in Table 1) are excluded one at a time. It can be noted that lexical group of features were most important without which the performance drops significantly. MetaMap matches were the next most important group of features. Each of the remaining feature groups improves the performance by only small amount.

Features	P	R	F
All	0.829	0.673	0.743
All - Lexical	0.779	0.569	0.658
All - Semantic	0.824	0.669	0.738
All - MetaMap	0.810	0.648	0.720
All - Lemmatization	0.825	0.666	0.737
All - Abbreviations	0.828	0.668	0.740

Table 4: Ablation study results for Task A showing how the performance is affected by excluding various feature groups (shown in Table 1). Development data was used for testing. Only strict precision (P), recall (R) and F-measure (F) are shown.

Component	Accuracy
Training	78.1
UMLS	83.8
Training + UMLS	88.8
Training + Patterns	86.3
UMLS + Patterns	85.2
Training + UMLS + Patterns	89.5

Table 5: Performance on Task B obtained by combinations of exactly matching the mentions in the training data, exactly matching in the UMLS and using learned edit distance patterns for approximately matching in the UMLS. Development data was used for testing with gold-standard disease mentions.

Table 5 shows the performance on Task B when disease mentions are exactly matched in the training data, exactly matched in the UMLS and approximately matched in the UMLS using edit distance patterns, as well as their combinations. In order to evaluate the performance of our system on Task B independent of its performance on Task A, we used gold-standard disease mentions in the development data as input for Task B in which case the strict and relaxed accuracies are equal. It may be noted that adding edit distance patterns improves the performance in each case.

4 Conclusions

We participated in the SemEval 2014 Task 7 of disorder mention extraction and normalization from clinical text. Our system used conditional random fields as the learning method for the extraction task with various lexical, semantic and MetaMap based features. We introduced a new method to do approximate matching for normalization that learns general patterns of variations using edit distances. Our system performed competitively on both the tasks.

References

Alan R Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Mikhail Bilenko and Raymond J Mooney. 2003. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 39–48. ACM.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of 18th International Conference on Machine Learning (ICML-2001)*, pages 282–289, Williamstown, MA.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, page 707.

Naoaki Okazaki. 2007. CRFsuite: A fast implementation of Conditional Random Fields (CRFs), http://www.chokkan.org/software/crfsuite/.

Sameer Pradhan, Noemie Elhadad, B South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, W Chapman, and Guergana Savova. 2013. Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. *Online Working Notes of CLEF, CLEF*, 230.

Hua Xu, Peter D Stetson, and Carol Friedman. 2012. Combining corpus-derived sense profiles with estimated frequency information to disambiguate clinical abbreviations. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1004. American Medical Informatics Association.