IITP: A Supervised Approach for Disorder Mention Detection and Disambiguation

Utpal Kumar Sikdar, Asif Ekbal and Sriparna Saha
Department of Computer Science and Engineering
Indian Institute of Technology Patna, India
{utpal.sikdar,asif,sriparna}@iitp.ac.in

Abstract

In this paper we briefly describe our supervised machine learning approach for disorder mention detection system that we submitted as part of our participation in the SemEval-2014 Shared task. The main goal of this task is to build a system that automatically identifies mentions of clinical conditions from the clinical texts. The main challenge lies due in the fact that the same mention of concept may be represented in many surface forms. We develop the system based on the supervised machine learning algorithms, namely Conditional Random Field and Support Vector Machine. One appealing characteristics of our system is that most of the features for learning are extracted automatically from the given training or test datasets without using deep domain specific resources and/or tools. We submitted three runs, and best performing system is based on Conditional Random Field. For task A, it shows the precision, recall and F-measure values of 50.00%, 47.90% and 48.90%, respectively under the strict matching criterion. When the matching criterion is relaxed, it shows the precision, recall and F-measure of 81.50%, 79.70% and 80.60%, respectively. For task B, we obtain the accuracies of 33.30% and 69.60% for the relaxed and strict matches, respectively.

1 Introduction

The SemEval-2014 Shared Task 7 is concerned with the analysis of clinical texts, particularly for disorder mention detection and disambiguation. The purpose of this task is to enhance current research in Natural Language Processing (NLP) methods used in the clinical domain. The task is a continuation of the CLEF/eHealth ShARe 2013 Shared Task. In particular there were two specific tasks, *viz.* (i). **Task A:** To identify disorder mentions from biomedicine domain and (ii) **Task B:** To classify each mention with respect to the Unified Medical Language System (UMLS) Concept Unique Identifier (CUI). The task is challenging in the sense that the same mention of concept may be represented in many surface forms and mention may appear in the different parts of texts. Some systems (Cogley et al., 2013; Zuccon et al., 2013; Tang et al., 2013; Cogley et al., 2013) are available for disorder mention detection. Looking at the challenges and resources available at our hand we planned to adapt our existing system (Sikdar et al., 2013) for disorder mention detection. The original architecture was conceptualized as part of our participation in the BioCreative-IV Track-2 Shared Task on Chemical Compound and Drug Name Recognition. Although our submitted system for SemEval-14 shared task is in line with BioCreative-IV[1], it has many different features and characteristics.

We develop three systems (e.g. **Model-1**: sikdar.run-0, **Model-2**: sikdar.run-1 and **Model-3**: sikdar.run-2) based on the popular supervised machine learning algorithms, namely Conditional Random Field (CRF) (Lafferty et al., 2001) and Support Vector Machine (SVM) (Cortes and Vapnik, 1995; Joachims, 1999). The models were developed by varying the features and feature templates. A baseline model is constructed by using the UMLS MetaMap[2] tool. During testing we merge the development set with the training set. Evaluation results on test data with the benchmark set up show the F-measure values of

[1] www.biocreative.org/tasks/biocreative-iv/chemdner/

[2] http://mmtx.nlm.nih.gov/

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 314–318,
Dublin, Ireland, August 23-24, 2014.

48.90%, 46.50% and 46.50%, respectively under the strict criterion. Under relaxed matching criterion the models show the F-measure values of 80.60%, 78.20% and 79.60%, respectively. Our submission for Task-B is simple in nature where we consider only those mentions that are also predicted in the baseline model, i.e. only the common CUIs are considered. It shows the accuracies of 33.30%, 31.90% and 33.20%, respectively under strict matching criterion; and 69.60%, 69.60% and 69.10%, respectively under the relaxed matching criterion.

2 Method

Our method for disorder mention detection from clinical text is based on the supervised machine learning algorithms, namely CRF and SVM. The key focus was to develop a system that could be easily adapted to other domains and applications. We submitted three runs defined as below:
Model-1:sikdar.run-0: This is based on CRF, and makes use of the features as mentioned below.
Model-2:sikdar.run-1: This model is built by training a SVM classifier with the same set of features as CRF.
Model-3:sikdar.run-2: This model is constructed by defining a heuristics that looks at the outputs of both the models. For given instance, if one of the models predicts it to belong to the category of candidate disorder mention then this is given more priority in assigning the class. We observed performance improvement on the development set with this heuristic.

We identify and implement different features, mostly without using any deep domain knowledge or domain-specific external resources and/or tools. The features that are used to train the classifiers are briefly described below:

- **Context words:** Surrounding words carry effective information to identify disorder mention. In our case we consider the previous three and next three words as the features.

- **MetaMap match:** MetaMap is a widely used tool that maps biomedical mention to the UMLS CUI[3]. In UMLS, there are 11 semantic types denoting disorders. These are *Congenital Abnormality*, *Acquired Abnormality*, *Injury or Poisoning*, *Pathologic Function*, *Disease or Syndrome*, *Mental or Behavioral Dysfunction*, *Cell or Molecular Dysfunction*, *Experimental Model of Disease*, *Anatomical Abnormality*, *Neoplastic Process* and *Signs and Symptoms*. The training set is passed through the MetaMap, and then we prepare a list of mentions that belong to the UMLS semantic types. A feature is thereafter defined that takes a value of 1 if the current token appears in the list; otherwise the value becomes 0.

- **Part-of-Speech (PoS) Information:** In this work, we use PoS information of the current token as the feature. PoS information was extracted from the GENIA tagger[4] V2.0.2, which is a freely available resource.

- **Root words:** Stems or root words, which are extracted form GENIA tagger V2.0.2, are used as the feature.

- **Chunk information:** We use GENIA tagger V2.0.2 to extract the chunk information. It helps to identify the boundaries of disorder mentions.

- **Initial capital:** The feature is set to true if the first character of the current token is a capital letter.

- **All capital:** The feature is set to true if all the letters of the current token are capitalized.

- **Stop words:** A feature is defined that is set to one if the current token appears in the list of stop words.

- **Word normalization:** Word shapes refer to the mapping of each word to their equivalence classes. Each capitalized character of the word is replaced by 'A', small characters are replaced by 'a' and digits are replaced by '0'.

- **Word suffix and prefix:** These features indicate the fixed-length character sequences (here 4) stripped either from the end (suffix) or beginning positions of words. This is useful in the sense that disorder mentions share some common sub-strings.

[3] http://www.nlm.nih.gov/research/umls/

[4] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger

- **Unknown word:** This feature is implemented depending upon whether the current token was found during training or not. For the training set this has been set randomly.

- **Word length:** If the length of a token is more than a predefined threshold (here 5) then it is most likely a disorder mention. This feature is defined with the observation that very short words are most probably not disorder mentions.

- **Alpha digit:** If the current token contains digit character(s), then the feature is set to true otherwise false.

- **Informative words**: This feature is developed from the training dataset. The words or the sequence of words that precede and follow the disorder mentions could be useful for mention detection. The most frequently occurring words that appear within the context of $w_{i-2}^{i+2} = w_{i-2} \dots w_{i+2}$ of w_i are extracted from the training data. Two different lists are prepared, one for the informative words that precede the mentions and the other contains the informative words that follow the mentions. Thereafter we define two features that fire for the words of these lists.

- **Disorder mention prefix and suffix**: We extract most frequently occurring prefixes and suffixes of length 2 from the disorder mentions present in the training data. We prepare two lists containing the prefix and suffix sub-sequences (of length two) that appear at least 10 times in the training set. We define two features that go on/off depending upon whether the current word contains any subsequence present in the lists.

- **Dynamic information**: The feature is extracted from the output label(s) of the previous token(s). The feature value is determined at run time.

3 Experimental Results

3.1 Datasets

In SemEval-2014 Shared task 7, three types of data were provided- training, development and test. Training data contains four different types of notes- discharge, ecg, echo and radiology. Development data consists of notes of three different domains, *viz.* discharge, echo and radiology. But the test set contains only the discharge notes. For a given document, the start and end indices are mentioned for the disorder mentions. There are 199, 99 and 133 documents in the training, development and test set, respectively.

3.2 Results and Analysis

We use a regular expression based simple pattern (e.g. dot and space) matching techniques for the sentence splitting and tokenization. We use C^{++} based CRF^{++} package[5] for CRF experiments. We set the default values of the following parameters (a). the hyper-parameter of CRF. With larger value, CRF tends to overfit to the given training data; (b). parameter which sets the cut-off threshold for the features (default value is 1). CRF uses only those features, having more than the cut-off threshold in the given training data.

In case of SVM we used YamCha[6] toolkit along with TinySVM[7]. We use the polynomial kernel function of degree two. In order to denote the boundaries of a multi-word disorder mention properly we use the standard BIO encoding scheme, where B, I and O represent the beginning, intermediate and outside, respectively, for a multi-word token. Please note that the mentions are not continuous, i.e. they could appear at the various positions of the text. For example, in the sentence *The left atrium is moderately dilated*, there is a single mention *left atrium dilated.* Its BIO format is represented in Table 1.

Token	**Tag**
The	O
left	B-Men
atrium	I-Men
is	O
moderately	O
dilated	I-Men
.	O

Table 1: An example of BIO representation.

Experiments are conducted on the benchmark setup as provided by the competition organizer. At first we train our system using the training set and evaluate using the development set in order to de-

[5] http://crfpp.sourceforge.net

[6] http://chasen-org/ taku/software/yamcha/

[7] http://chasen.org/ taku/software/TinySVM/

System	Strict			Relaxed		
	P	R	F	P	R	F
Baseline	19.9	29.0	23.6	44.9	63.0	52.4
Model-1	52.5	43.0	47.3	86.2	72.6	78.8
Model-2	49.3	41.0	44.8	82.8	70.6	76.2
Model-3	46.7	44.0	45.3	81.2	77.5	79.3

Table 2: Results on development set for Task A.

System	Strict	Relaxed
	Accuracy	Accuracy
Baseline	24.6	85.1
Model-1	31.2	72.5
Model-2	29.9	73.0
Model-3	31.8	72.4

Table 3: Results on development set for Task B.

termine the best configuration. We define a baseline model by passing the development set to the UMLS MetaMap tool. Its results along with the baseline model are reported in Table 2 for Task A. Evaluation shows that our proposed system performs reasonably better compared to the baseline model. It is also to be noted that Model-1 performs better compared to the other two submitted models for the strict matching, but for relaxed evaluation, Model-3 performs better than Model-1 and Model-2. Under strict matching criterion, Model-1 achieves 2.7% and 5.0% increments in precision over the second and third models, respectively. For relaxed matching, Model-3 achieves 4.9% and 6.9% increments in recall over the first and second models, respectively. Results on the development set for Task-B are reported in Table 3. Please note that although our system performs better than the baseline in terms of strict matching, it does not show better accuracy under relaxed matching criterion. This is because our system for Task-B is developed by considering only those mentions that lie in the intersection of baseline and CRF models. As a result many mentions are missed. During final submissions we merged development sets with the respective training sets, and perform evaluation on the test sets. We report our results on the test sets in Table 4 and Table 5 for Task-A and Task-B, respectively.

System	Strict			Relaxed		
	P	R	F	P	R	F
Model-1	50.0	47.9	48.9	81.5	79.7	80.6
Model-2	47.3	45.8	46.5	78.9	77.6	78.2
Model-3	45.0	48.1	46.5	76.9	82.6	79.6

Table 4: Evaluation results on test set for Task A.

System	Strict	Relaxed
	Accuracy	Accuracy
Model-1	33.3	69.6
Model-2	31.9	69.6
Model-3	33.2	69.1

Table 5: Results of Task B for the test set.

We carefully analyze the results and find that most of the errors encountered because of the discontiguous mentions. Different components of a mention may be mapped to the different concepts. In our system we treat two mentions as a single unit if they have some shared tokens. For example, the sentence “She also notes new sharp pain in left shoulder blade/back area” contains two different mentions, *viz.*“pain shoulder blade” and “pain back”. Here shared word of these two mentions is “pain”, but we consider these two mentions as a single unit such as “pain shoulder blade back”. This contributes largely to the errors that our system faces for the first task. For the second task, we miss a number of mentions, and this can be captured if we directly match the system identified mentions to the entire UMLS database.

4 Conclusion

In this paper we report on our works as part of our participation in the SemEval-2014 shared task related to clinical text mining. We submitted three runs for both the tasks, *viz.* disorder mention detection and disambiguation. Our submitted runs for the first task are based on CRF and SVM. We make use of a set of features that are not very domain-specific. The system developed for the second task is very simple and is based on UMLS Meta Map tool.

There are many avenues for future research: identification of more features for the first task; use of some domain-specific resources and/or tools for the first task; use of entire UMLS thesaurus for mapping the disorder mentions; use of some machine learning techniques for disambiguation. We also plan to investigate how systematic feature selection, ensemble learning and machine learning optimization have impact on disorder mention detection and disambiguation.

References

James Cogley, Nicola Stokes, and Joe Carthy. 2013. Medical Disorder Recognition with Structural Support Vector Machines. *In Proceedings of CLEF*.

Corinna Cortes and Vladimir Vapnik. 1995. Support Vector Networks. *Machine Learning*, 20:273–297.

Thorsten Joachims, 1999. *Making Large Scale SVM Learning Practical*, pages 169–184. MIT Press, Cambridge, MA, USA.

John Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML*, pages 282–289.

Utpal Kumar Sikdar, Asif Ekbal, and Sriparna Saha. 2013. Domain-independent Model for Chemical Compound and Drug Name Recognition. *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop*, vol. 2:158–161.

Buzhou Tang, Yonghui Wu, M. Jiang, J. C. Denny, and Hua Xu. 2013. Recognizing and Encoding Disorder Concepts in Clinical Text using Machine Learning and Vector Space Model. *In Proceedings of CLEF*.

Guido Zuccon, A. Holloway, B. Koopman, and A. Nguyen. 2013. Identify Disorders in Health Records using Conditional Random Fields and Metamap. *In Proceedings of CLEF*.