Conditional Random Fields and Support Vector Machines for Disorder Named Entity Recognition in Clinical Texts

Dingcheng Li
University of Minnesota
Minneapolis, Minnesota, USA
lixxx345@umn.edu

Karin Kipper-Schuler
Mayo Clinic College of Medicine
Rochester, Minnesota, USA
schuler.karin@mayo.edu

Guergana Savova
Mayo Clinic College of Medicine
Rochester, Minnesota, USA
savova.guergana@mayo.edu

Abstract

We present a comparative study between two machine learning methods, Conditional Random Fields and Support Vector Machines for clinical named entity recognition. We explore their applicability to clinical domain. Evaluation against a set of gold standard named entities shows that CRFs outperform SVMs. The best F-score with CRFs is 0.86 and for the SVMs is 0.64 as compared to a baseline of 0.60.

1 Introduction and background

Named entity recognition (NER) is the discovery of named entities (NEs), or textual mentions that belong to the same semantic class. In the biomedical domain NEs are diseases, signs/symptoms, anatomical signs, and drugs. NER performance is high as applied to scholarly text and newswire narratives (Leaman et al., 2008). Clinical free-text, on the other hand, exhibits characteristics of both informal and formal linguistic styles which, in turn, poses challenges for clinical NER. Conditional Random Fields (CRFs) (Lafferty et al., 2001) and and Support Vector Machines (SVMs) (Cortes and Vapnik, 1995) are machine learning techniques which can handle multiple features during learning. CRFs' main strength lies in their ability to include various unrelated features, while SVMs' in the inclusion of overlapping features. Our goal is to compare CRFs and SVMs performance for clinical NER with focus on disease/disorder NEs.

2 Dataset and features

Our dataset is a gold standard corpus of 1557 single- and multi-word disorder annotations (Ogren et al., 2008). For training and testing the CRF and SVM models the IOB (inside-outside-begin) notation (Leaman, 2008) was applied. In our project, we used 1265 gold standard annotations for training and 292 for testing. The features used for the learning process are described as follows. *Dictionary look-up* is a binary value feature that represents if the NE is in the dictionary (SNOMED-CT). *Bag of Words (BOW)* is a representation of the context by the unique words in it. *Part-of-speech tags (POS)* of BOW is the pos tags of the context words. *Window size* is the number of tokens representing context surrounding the target word. *Orientation(*left or right) is the location of the feature in regard to the target word. *Distance* is the proximity of the feature in regard to the target word *Capitalization* has one of the four token-based values: all upper case, all lower case, mixed_case and initial upper case. *Number features* refer to the presence or absence of related numbers. Feature sets are in Table 1.

3 Results and discussion

Figure 1 shows the CRF results. The F-scores, recall and precision for the baseline dictionary look-up are 0.604, 0.468 and 0.852 respectively. When BOW is applied in feature combination 2 results improve sharply adding 0.15, 0.17 and 0.08 points respectively. The F-score, recall and precision improve even further with the capitalization feature to 0.858, 0.774 and 0.963 respectively. Figure 2 shows SVM results. The addition of more features to the model did not show an upward trend. The best results are with feature combination 1 and 3. The F-score reaches 0.643, which although an improvement over the baseline greatly underperforms CRF results. BOW features seem not discriminative with SVMs. When the window size increases to 5, performance decreases as demonstrated in feature combinations 2, 4 and 8. Results with feature combination 4, in particular, has a pronounced downward trend. Its F-score is 0.612, a decrease by 0.031 compared with Test 1 or Test 3. Its recall and precision are 0.487 and 0.822 respectively, a decrease by 0.036 and 0.01 respectively. This supports the results achieved with CRFs where a smaller window size yields better performance.

BioNLP 2008: Current Trends in Biomedical Natural Language Processing, pages 94–95,
Columbus, Ohio, USA, June 2008. ©2008 Association for Computational Linguistics

No	*Features*
1	dictionary look-up (baseline)
2	dictionary look-up+BOW+Orientation+distance (Window 5)
3	dictionary look-up + BOW + Orientation + distance (Window 3)
4	dictionary look-up + BOW + POS + Orientation + distance (Window 5)
5	dictionary look-up + BOW +POS + Orientation + distance (Window 3)
6	dictionary look-up + BOW +POS + Orientation + distance (Window 3) + bullet number
7	dictionary look-up + BOW + POS + Orientation + distance(Window 3) + measurement
8	dictionary look-up + BOW + POS + Orientation + distance (Window 5) + neighboring number
9	dictionary look-up + BOW +POS + Orientation + distance (Window 3) + neighboring number
10	dictionary look-up + BOW +POS + Orientation + distance (Window 3)+neighboring number+measurement
11	dictionary look-up+BOW+POS+Orientation (Window 3)+neighboring number+bullet number + measurement
12	dictionary look-up + BOW +POS + Orientation +distance (Window 3) + neighboring number + bullet number + measurement + capitalization

Table 1: Feature combinations

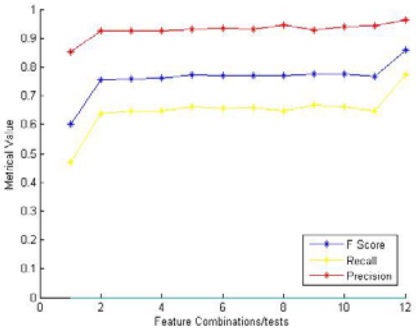

Figure 1: CRF evaluation results

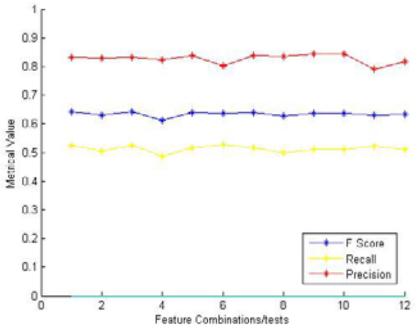

Figure 2: SVM evaluation results

As the results show, context represented by the BOW feature plays an important role indicating the importance of the words surrounding NEs. On the other hand, POS tag features did not bring much improvement, which perhaps hints at a hypothesis that grammatical roles are not as important as context in clinical text. Thirdly, a small window size is more discriminative. Clinical notes are unstructured free text with short sentences. If a larger window size is used, many words will share similar features. Fourthly, capitalization is highly discriminative. Fifthly, as a finite state machine derived from HMMs, CRFs can naturally consider state-to-state dependences and feature-to-state dependences. On the other hand, SVMs do not consider such dependencies. SVMs separate the data into categories via a kernel function. They implement this by mapping the data points onto an optimal linear separating hyperplane. Finally, SVMs do not behave well for large number of feature values. For large number of feature values, it would be more difficult to find discriminative lines to categorize the labels.

4 Conclusion and future work

We investigated the use of CRFs and SVMs for disorder NER in clinical free-text. Our results show that, in general, CRFs outperformed SVMs. We demonstrated that well-chosen features along with dictionary-based features tend to improve the CRF model's performance but not the SVM's.

Acknowledgements

The work was partially supported by a Biomedical Informatics and Computational Biology scholarship from the University of Minnesota.

References

Corinna Cortes and Vladimir Vapnik. Support-vector network. *Machine Learning*, 20:273-297, 1995.

John Lafferty, Andrew McCallum and Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning* (ICML-2001), 2001.

Robert Leaman and Graciela Gonzalez. BANNER: an Executable Survey of Advances in Biomedical Named Entity Recognition. *Pacific Symposium on Biocomputing* 13:652-663. 2008.

Philip Ogren, Guergana Savova and Christopher G Chute. Constructing evaluation corpora for automated clinical named entity recognition. *Proc LREC* 2008.