BioCreative II Gene Mention Tagging System at IBM Watson

Rie Kubota Ando
rie1@us.ibm.com
IBM T.J. Watson Research Center, 19 Skyline Drive, Hawthorne, New York 10532, USA

Abstract

This paper describes our system developed for the BioCreative II gene mention tagging task. The goal of this task is to annotate mentions of genes or gene products in the given Medline sentences. Our focus was to experiment with a semi-supervised learning method, *Alternating Structure Optimization (ASO)* [1], by which we exploited a large amount of *unlabeled data* in addition to the labeled training data provided by the organizer. The system is also equipped with automatic induction of high-order features, gene name lexicon lookup, classifier combination, and simple post-processing. Our system appears to be competitive. All of our three official runs belong to the Quartile 1.

1 Gene mention tagging system

Our gene mention tagging system was built on top of a named entity chunking system described in [1], which was used for annotating names of persons, organizations, and so forth. This system casts the chunking task into that of sequential labeling, as is commonly done, by encoding chunk information into token tags. It uses a regularized linear classifier with modified Huber loss and the 2-norm regularization. That is, using the 'one-versus-all' scheme, we train binary classifiers, one for each token tag, using n labeled data points $\{(\mathbf{x}_i, y_i)\}$ for $i = 1, \ldots, n$ by: $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} L(\mathbf{w}^T\mathbf{x}_i, y_i) + \lambda||\mathbf{w}||^2$. The regularization parameter λ is set to 10^{-4}. L is the loss function: $L(p, y) = \max(0, 1 - py)^2$ if $py \geq -1$; and $-4py$ otherwise. The optimization is done by stochastic gradient descent. Viterbi-style dynamic programming is performed to find the token tag sequence with the largest confidence. Feature types are shown in Figure 1. Using this framework, we experimented with additional resources and algorithms, which we describe below.

· words, parts-of-speech, character types, 4 characters at beginning/ending in a 5-word window
· words in a 3-syntactic chunk window.
· labels assigned to two words on the left.
· bi-grams of the current word and the left label.
· labels assigned to previous occurrences of the current word.

Figure 1: Feature types.

1.1 Exploiting unlabeled data through Alternating Structure Optimization (ASO)

ASO is a multi-task learning algorithm that seeks to improve performance on individual tasks by simultaneously learning multiple tasks that are related to each other. The application of ASO to semi-supervised learning involves automatic generation of thousands of prediction problems (called 'auxiliary problems') and their labeled data from unlabeled data, so that the multi-task learning algorithm can be applied on the unlabeled data.

To put this into perspective, ASO-based semi-supervised learning can be viewed as learning new (and better) feature representation from unlabeled data. This is done by learning auxiliary predictors that predict one part of the feature vectors from another part of the feature vectors, which can be learned from unlabeled data. Under certain conditions, it can be shown that learning auxiliary predictors of this type can reveal the predictive structure (something useful for the target prediction problems) underlying the data. The final classifiers are trained with labeled data using the original features and the new features learned from unlabeled data. Since modern classifiers based on empirical risk minimization are capable of ignoring irrelevant features to some degree, the risk of using unlabeled data this way is relatively low, and its potential gain is large. [1] should be consulted for the details of ASO. Below, we only describe the specifics of our setting.

Auxiliary problems The 'word prediction' auxiliary problems were used with the same implementation details as in [1].

Unlabeled data For unlabeled data, we had about 5 million Medline abstracts (consisting of approx. 500 million words) over a 10-year period (1994–2003) at hand. To utilize the entire corpus through ASO, the training of thousands of predictors on these 500 million words were required, which we felt would be too resource-intensive. However, our experiments using the old BioCreative I data set indicated that if we use a small subset of the unlabeled corpus generated by random sentence selection, the performance improvement from ASO was marginal. This was due to the small size of the vocabulary overlap between the unlabeled data and the training/test data. (This issue appears to be specific to the biomedical domain, which has a much larger vocabulary than, for instance, the news domain.) To benefit from unlabeled data with reasonable computation time, we created a small but useful unlabeled corpus as follows. We go through every sentence of the input corpus while counting up the occurrences of words. If the sentence contains at least one word that has occurred less than k times so far, then we choose this sentence; discard it otherwise. By setting $k = 25$, we obtained an unlabeled corpus that is much smaller than the original one but represents well (to some degree) the entire vocabulary of the original corpus.

1.2 Automatic induction of high-order features

High-order features (combining two or more base features) are sometimes effective, but generating all the combinations would make training expensive. We used a simple method for selectively generating bi-gram features. The idea is to select a bi-gram feature only if it would help to correctly classify the data point that was misclassified when only base features (in Figure 1) were used. This is done as follows. First, we train a classifier using the base features only on a labeled data set L_1. Next, we test this classifier on a labeled data set L_2. We generate bi-gram features (e.g., 'current-word="gene" and next-word="(" ') only from the data points that are misclassified. We filter out all but those occurring in at least q misclassified data points. To further filter out the bi-grams, we consider $2K$ criteria for K-way classification, each of which inspects whether that bi-gram is useful as evidence for being positive/negative with respect to each class. According to each criteria, each bi-gram receives a score computed as a sum of partial derivatives of the loss function on the respective data points. The bi-gram is selected if its score is within top t according to one of the $2K$ criteria. We set $q = 10$ and $t = 100000$. Although one could divide the training set into L_1 and L_2 disjointly, we instead used the entire training set as L_1 and L_2 and applied an 'early stop' when training the classifier with base features.

1.3 Domain lexicons

Our two (out of three) official runs used a domain lexicon, which we generated from LocusLink, Swiss-Prot, and Mesh. Our domain lexicon consists of a list of names (e.g., "adenosine arabinose") with tags that indicate the information source (e.g., "MESH"). In the feature generation process, we turn on the corresponding feature according to the tags associated with the matched name entries (including partial matching).

1.4 Classifier combination

A number of studies have shown that combining results of several classifiers (that, ideally, produce similar performance but make different mistakes) often improves performance over a single classifier. The classifiers to be combined could be, for instance, those employing different schemes of chunk encoding (e.g., one classifier uses BIO, and another uses EIO) or those based on different models (e.g., one is MaxEnt, and the other is SVM). [2] reported that combining a left-to-right chunker and a right-to-left chunker (by taking a union of the two sets of annotations) was effective on the BioCreative I data. We adopt this strategy and combine the results of a left-to-right chunker and a right-to-left chunker. However, instead of taking a union, we remove any annotation that overlaps with another by keeping longer ones, which performed better than taking a union in our experiments on the BioCreative I data.

	Post-processing	Feature induction	Name lexicons	Classifier combination	Unlabeled data	P	R	F	
Baseline	–	–	–	–	–	89.13	79.39	83.98	–
Post-processing	X	–	–	–	–	89.40	79.39	84.10	(+0.12)
Feature induction	–	X	–	–	–	89.11	79.86	84.23	(+0.25)
Name lexicon	–	–	X	–	–	88.89	80.48	84.47	(+0.49)
Classifier combination	–	–	–	X	–	85.14	84.90	85.02	(+1.04)
Unlabeled data	–	–	–	–	X	91.17	81.52	86.07	(+2.09)
Run#3	X	X	X	–	X	**91.54**	81.99	86.50	(+2.52)
Run#1	X	X	–	X	X	88.37	85.94	87.14	(+3.16)
Run#2	X	X	X	X	X	88.48	**85.97**	**87.21**	(+3.23)

Figure 2: Performance results. Effectiveness of the five components; the three official runs. The best performance in each column is highlighted. The numbers in parentheses are performance improvements over the baseline (a supervised configuration using base features).

1.5 Simple post-processing

Many of the BioCreative I systems were equipped with some post-processing. We adopt the one used in [2], which removes annotations that include any unmatched parenthesis.

2 Performance results

Figure 2 shows the performance of our official runs and some post-submission experimental results. ‘Baseline’ is a standard supervised configuration using the features in Figure 1. The five rows following ‘Baseline’ compare the performance improvements obtained by the five components described above. The performance trend is consistent with that of our experiments on the BioCreative I data (not included in this paper). All the five components improved performance on the BioCreative II evaluation data as well as the BioCreative I data. The largest performance gain is obtained from unlabeled data through ASO. This semi-supervised configuration (‘Unlabeled data’) achieves 2.09 higher F-measure than the baseline. Also note that it outperforms the baseline both in precision and recall. The second best contributor is classifier combination, which improved recall at the price of precision and resulted in 1.04 improvement in F-measure.

Among the three official runs, Run#2 that uses all the five components achieved the best performance (3.23 higher than our baseline) among all of our configurations. These results clearly confirm the effectiveness of our approach on this task.

3 Conclusion

This paper presented the gene mention tagging system that participated in BioCreative II. The main strength of the system derives from semi-supervised learning using the ASO algorithm. We also experimented with classifier combination, domain lexicon, automatic generation of high-order features, and simple post-processing, which were all effective.

Since our approach is general, we expect it to be also useful for tagging other types of mentions in the biomedical text. We presume that semi-supervised learning is particularly suitable for exploring biomedical texts, given the presence of a huge amount of unlabeled data – the Medline corpus.

References

[1] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

[2] S. Dingare, J. Finkel, C. D. Manning, M. Nissim, and B. Alex. Exploring the boundaries: Gene and protein identification in biomedical text. In *Proceedings of the BioCreative Workshop*, 2004.