Boosting the protein name recognition performance by bootstrapping on selected text

Yue Wang and **Jin-Dong Kim**
Database Center for Life Science,
Research Organization of Information and Systems
2-11-16 Yayoi, Bunkyo-ku, Tokyo, Japan 113-0032
{wang,jdkim}@dbcls.rois.ac.jp

Abstract

When only a small amount of manually annotated data is available, application of a bootstrapping method is often considered to compensate for the lack of sufficient training material for a machine-learning method. The paper reports a series of experimental results of bootstrapping for protein name recognition. The results show that the performance changes significantly according to the choice of text collection where the training samples to bootstrap, and that an improvement can be obtained only with a well chosen text collection.

1 Introduction

While machine learning-based approaches are becoming more and more popular for the development of natural language processing (NLP) systems, corpora with annotation are regarded as a critical resource for the training process. Nonetheless, the creation of corpus annotation is an expensive and time-consuming work (Cohen et al., 2005), and it is often the case that lack of sufficient annotation hinders the development of NLP systems. Bootstrapping method (Becker et al., 2005; Vlachos and Gasperin, 2006) can be considered as a way to automatically inflate the amount of corpus annotation to complement the lack of sufficient annotation.

In this study, we report the experimental results on the effect of bootstrapping for the training of protein name recognizers, particularly in the situation when we have only a small amount of corpus annotations.

In summary, we begin with a small corpus with manual annotation for protein names. A named entity tagger trained on the small corpus is applied to a big collection of text, to obtain more annotation. We hope the newly created annotation to be precise enough so that the training of a protein tagger can benefit from the increased training material.

We assume that the accuracy of a bootstrapping method (Ng, 2004) depends on two factors: the accuracy of the bootstrap tagger itself and the similarity of the text to the original corpus. While accuracy of the bootstrap tagger may be maximized by finding the optimal parameters of the applied machine learning method, the choice of text where the original annotations will bootstrap may also be a critical factor for the success of the bootstrapping method.

Experimental results presented in this paper confirm that we can get a improvement by using a bootstrapping method with a well chosen collection of texts.

The paper is organized as follows. Section 2 introduces the two datasets used in this paper. Following that, in Section 3, we briefly introduce the experiments performed in our research. The experimental results are demonstrated in Section 4. The research is concluded in Section 5 and in the meanwhile, future work is discussed.

2 Datasets

2.1 The cyanobacteria genome database

Cyanobacteria are prokaryotic organisms that have served as important model organisms for studying oxygenic photosynthesis and have played a signifi-

Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), pages 240–243,
Montréal, Canada, June 8, 2012.

cant role in the Earthfs history as primary producers of atmospheric oxygen (Nakao et al., 2010).

The cyanobacteria genome database (abbreviated to CyanoBase[1]) includes the annotations to the PubMed text. In total, 39 species of the cyanobacteria are covered in the CyanoBase.

In our cyanobacteria data (henceforth, the Kazusa data for short), 270 abstracts were annotated by two independent annotators. We take the entities, about which both of the annotators agreed with each other. In total, there are 1,101 entities in 2,630 sentences.

The Kazusa data was split equally into three subsets and the subsets were used in turn as the training, development and testing sets in the experiments.

2.2 The BioCreative data

The BioCreative data, which was used for the BioCreative II gene mention task[2], is described as the tagged gene/protein names in the PubMed text. The training set is used in the research, and totally there are 15,000 sentences in the dataset.

Unlike other datasets, the BioCreative data was designed to contain sentences both with and without protein names, in a variety of contexts. Since the collection is made to explicitly compile positive and negative examples for protein recognition, there is a chance that the sample of text is not comprehensive, and gray-zone expressions may be missed.

The reason that we chose the BioCreative data for the bootstrapping is that, the BioCreative data (henceforth, the BC2 data for short) is the collection for the purpose of training and evaluation of protein name taggers.

3 Experiment summary

In the following experiments, the NERSuite[3], a named entity tagger based on Conditional Random Fields (CRFs) (Lafferty et al., 2001; Sutton and McCallum, 2007), is used. The NERSuite is executable open-source and serves as a machine learning system for named entity recognition (NER). The sigma value for the L_2-regularization is optimizable and in our experiments, we tune the sigma value between 10^{-1} to 10^4.

[1]http://genome.kazusa.or.jp/cyanobase

[2]http://www.biocreative.org/

[3]http://nersuite.nlplab.org/

As mentioned in Section 2.1, the three subsets of Kazusa data are used for training, tuning and testing purposes, in turn. We experimented with all the six combinations.

Experiments were performed to compare three different strategies. First, with the *baseline strategy*, the protein tagger is trained only on the Kazusa training set. The sigma value is optimized on the tuning set, and the performance is evaluated on the test set. It is the most typical strategy particularly when it is believed there is a sufficient training material.

Second, with the *bootstrapping strategy*, the Kazusa training set is used as the seed data. A tagger for bootstrapping (bootstrap tagger, hereafter) is trained on the seed data, and applied to the BC2 data to bootstrap the training examples. Another protein tagger (application tagger) is then trained on the bootstrapped BC2 data together with the seed data. The Kazusa tuning set is used to optimize the two sigma values for the two protein taggers, and the performance is evaluated on the test set. With this strategy, we wish the bootstrapped examples complement the lack of sufficient training examples.

Experiment	Seed	BT	BT+SS
E1	368	647	647 (1,103)
E2	368	647	647 (1,103)
E3	366	759	759 (1,200)
E4	366	769	590 (1,056)
E5	367	882	558 (1,068)
E6	367	558	558 (1,068)

Table 1: The number of positive examples used in each experiment. The "BT" column shows the number of positive examples obtained by the bootstrapping in the 15,000 BC2 sentences. In the last column, the figures in parentheses are the number of the selected sentences.

Third, the *bootstrapping with sentence selection strategy* is almost the same with the bootstrapping strategy, except that the second tagger is trained after the non-relevant sentences are filtered out from the BC2 data. Here, non-relevant sentences mean those that are not tagged by the the bootstrap tagger. With this strategy, we wish an improvement with the bootstrapping by removing noisy data. Table 1 shows the number of the seed and bootstrapped examples used for the three strategies. It is observed that the seed

	Training	Tuning	Testing	Baseline	BT	BT+SS
E1	A	B	C	63.7/29.2/40.0 [10^2]	61.3/25.9/36.4 [10^4-10^1]	61.7/38.2/47.1 [10^4-10^4]
E2	A	C	B	65.2/36.9/47.1 [10^3]	67.7/35.0/46.1 [10^4-10^1]	61.7/46.7/53.2 [10^4-10^4]
E3	B	C	A	75.3/36.4/49.1 [10^2]	75.2/31.3/44.2 [10^2-10^1]	67.1/40.0/50.1 [10^2-10^1]
E4	B	A	C	68.5/33.8/45.3 [10^2]	70.2/28.9/40.9 [10^4-10^1]	66.7/36.5/47.2 [10^1-10^2]
E5	C	B	A	77.7/35.1/48.3 [10^1]	71.8/27.7/40.0 [10^4-10^2]	70.9/38.3/49.7 [10^0-10^1]
E6	C	A	B	73.0/39.1/50.9 [10^1]	76.1/32.2/45.3 [10^0-10^2]	67.7/41.8/51.7 [10^0-10^2]

Table 2: Experimental results of using the Kazusa and BC2 data (Precision/Recall/F-score). "BT" and "SS" represent the bootstrapping and sentence selection strategies, respectively. The figures in square brackets are the sigma values optimized in the experiments.

annotation bootstrap only on a small portion of the BC2 data set, e.g., 1,103 vs. 15,000 sentences in the case of E1 (less than 10%), suggesting that a large portion of the data set may be irrelevant to the original data set.

4 Experimental results

The experimental results of all the six combinations are shown in Table 2. The use of the three subsets, denoted by A, B, C, of the Kazusa data set for training, tuning and testing in each experiment is specified in "training", "tuning" and "testing" columns. The results of the baseline strategy that uses only the Kazusa data are shown in the "baseline" column, whereas the results with the bootstrapping methods with and without sentence selection are shown in the last two columns. As explained in Section 3, the sigma values are optimized using the tuning set for each experiment. Note that for bootstrapping, we need two sigma values for the bootstrapping tagger and the application tagger. See section 3.

The performance of named entity recognition is measured in terms of precision, recall and F-score. For matching criterion, in order to avoid underestimation, instead of the exact matching, system performance is evaluated under a soft matching, the overlapping matching criterion. That is, if any part of the annotated protein/gene names is recognized by the NER tagger, we will regard that as a correct answer.

4.1 Results with the bootstrapping strategy

Comparing the two columns, "baseline" and "BT", we observe that the use of bootstrapping may lead to a degradation of the performance. Note that the sigma values are optimized on the development set for each experiment, and the text for bootstrapping is BC2 corpus which is expected to be similar to the Kazusa corpus, but still it is observed that the bootstrapping does not work, suggesting that the text collection may not yet similar enough.

4.2 Results with bootstrapping with sentence selection

Comparing the last column (the "BT+SS" column) to the "baseline" column, we observe that the application of the bootstrapping method with sentence selection consistently improves the performance. The improvement is sometimes significant, e.g., 7.1% of difference in F-score in the case of E1, but sometimes not, e.g., only 0.8% in the case of E6, but the performance is improved in the every experiments. The results confirm our assumption that the choice of text for bootstrapping is important, and that the sentence selection is a stable method for the choice of text.

5 Conclusion and future work

In order to compensate for the lack of sufficient training data for a CRF-based protein name recognizer, the potential of a bootstrapping method has been explored through a series of experiments. The BC2 data was chosen for the bootstrapping as the data set was one collected for protein name recognition.

Our initial experiment showed that the seed annotations bootstrapped only on a very small portion of the BC2 data set, suggesting that a big portion of the data set might be less relevant to the seed corpus. From a series of experiments, it was observed that the performance of protein name recognition was always improved with bootstrapping by selecting only

the sentences where the seed annotations bootstrap, and by using them as an additional training data.

The goal was to be able to predict more possible protein mentions (recall) at a relatively satisfactory level of the quality (precision). The experimental results suggest us, in order to achieve the goal, the choice of text collection is important for the success of the use of a bootstrapping method.

For the future work, we would like to take use of the original annotations in the BC2 data. A filtering strategy (Wang, 2010) will be performed. Instead of completely using the output of the Kazusa-trained tagger, we compare the output of the Kazusa-trained tagger with the BioCreative annotations. If the entity is recognized by the tagger and also annotated in the BioCreative data, then the annotation to this entity will be kept. The entity will be regarded as a true positive according to the BioCreative annotations. Otherwise, we will remove the annotation to the entity from the BioCreative annotations.

Further, we also would like to combine the bootstrapping with the filtering. Besides keeping the true positives, we also want to include some false positives from the bootstrapping. Because these false positives helps in improving the recall, when the tagger is applied to the Kazusa testing subset. To discriminate this strategy from the bootstrapping and filtering strategies, different sigma value should be used.

Acknowledgement

We thank Shinobu Okamoto for providing the Kazusa data and for many useful discussion. This work was supported by the "Integrated Database Project" funded by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan.

References

K. Bretonnel Cohen, Lynne Fox, Philip Ogren and Lawrence Hunter. 2005. Empirical data on corpus design and usage in biomedical natural language processing. *Proceedings of the AMIA Annual Symposium*, 38–45.

Markus Becker, Ben Hachey, Beatrice Alex, Claire Grover. 2005. Optimising Selective Sampling for Bootstrapping Named Entity Recognition. *Proceedings of the Workshop on Learning with Multiple Views*, 5–11.

Andreas Vlachos and Caroline Gasperin. 2006. Bootstrapping and Evaluating Named Entity Recognition in the Biomedical domain. *Proceedings of the BioNLP Workshop*, 138–145.

Andrew Ng. 2004. Feature selection, L_1 vs. L_2 regularization, and rotational invariance. *Proceedings of the 21st International Conference on Machine Learning (ICML)*.

Mitsuteru Nakao, Shinobu Okamoto, Mitsuyo Kohara, Tsunakazu Fujishiro, Takatomo Fujisawa, Shusei Sato, Satoshi Tabata, Takakazu Kaneko and Yasukazu Nakamura. 2010. CyanoBase: the cyanobacteria genome database update 2010. *Nucleic Acids Research*, 38:D379–D381.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the 18th International Conference on Machine Learning*, 282–289.

Charles Sutton and Andrew McCallum. 2007. An Introduction to Conditional Random Fields for Relational Learning. *Introduction to Statistical Relational Learning*, MIT Press.

Yue Wang. 2010. Developing Robust Protein Name Recognizers Based on a Comparative Analysis of Protein Annotations in Different Corpora. *University of Tokyo, Japan*, PhD Thesis.