Raising the Compatibility of Heterogeneous Annotations: A Case Study on Protein Mention Recognition

Yue Wang* **Kazuhiro Yoshida*** **Jin-Dong Kim*** **Rune Sætre*** **Jun'ichi Tsujii*†‡**
*Department of Computer Science, University of Tokyo
†School of Informatics, University of Manchester
‡National Center for Text Mining
Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033 JAPAN
{wangyue, kyoshida, jdkim, rune.saetre, tsujii}@is.s.u-tokyo.ac.jp

Abstract

While there are several corpora which claim to have annotations for protein references, the heterogeneity between the annotations is recognized as an obstacle to develop expensive resources in a synergistic way. Here we present a series of experimental results which show the differences of protein mention annotations made to two corpora, GENIA and AImed.

1 Introduction

There are several well-known corpora with protein mention annotations. It is a natural request to benefit from the existing annotations, but the heterogeneity of the annotations remains an obstacle. The heterogeneity is caused by different definitions of "protein", annotation conventions, and so on.

It is clear that by raising the compatibility of annotations, we can reduce the performance degradation caused by the heterogeneity of annotations.

In this work, we design several experiments to observe the effect of removing or relaxing the heterogeneity between the annotations in two corpora. The experimental results show that if we understand where the difference is, we can raise the compatibility of the heterogeneous annotations by removing the difference.

2 Corpora and protein mention recognizer

We used two corpora: the GENIA corpus (Kim et al., 2003), and the AImed corpus (Bunescu and Mooney, 2006). There are 2,000 MEDLINE abstracts and 93,293 entities in the GENIA corpus. The annotation is dependent on a small taxonomy of 36 classes. The AImed corpus consists of 225 MEDLINE abstracts, and there are 4,084 protein references.

Our protein mention recognizer is a Maximum Entropy Markov Model (MEMM) n-best tagger.

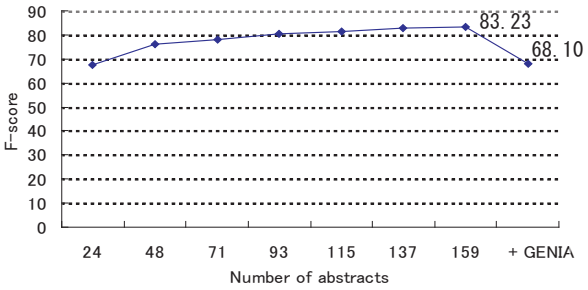

Figure 1: The learning curve according to the F-score

3 The effect of the inconsistency

We did two experiments in order to characterize the following two assumptions. First, we can improve the performance by increasing the size of the training data set. Secondly, the system performance will drop when more inconsistent annotations are introduced into the training data set.

In these two experiments, for the training, we used the AImed corpus and the AImed corpus plus the GENIA protein annotations, respectively. We conducted the evaluation on the AImed corpus.

The learning curve drawn from the results of the two mentioned experiments is shown in Figure 1. We can see that the learning curve is still increasing

BioNLP 2008: Current Trends in Biomedical Natural Language Processing, pages 118–119,
Columbus, Ohio, USA, June 2008.

Subcategory	Recall	Precision	F-score
Family_or_group	12.94	3.86	5.94
Domain_or_region	15.74	0.57	1.11
Molecule	48.80	34.43	40.37
Substructure	0.00	0.00	0.00
Subunit	65.36	3.38	6.43
Complex	13.43	0.98	1.83
ETC	14.29	0.03	0.07

Table 1: The experimental results on seven subclasses.

AImed + Subcategory	Criterion	F-score
Molecule+Subunit	Exact	64.72
	Left	69.48
	Right	67.64
Molecule+Subunit+Complex	Exact	63.76
	Left	72.77
	Right	67.60

Table 2: The experimental results on three subclasses.

when we used up all the training portions from the AImed corpus. Even though the rate of the improvement is slow, we would expect a further improvement if we could add more training data in a large scale, e.g. the GENIA corpus is 10 times bigger than the AImed corpus. But when we added the protein annotations in the GENIA corpus to the training data set, we witnessed a drastic degradation in the performance. We assume that the degradation is caused by the heterogeneity of the protein annotations in these two corpora, and we further assume that if the heterogeneity could be eliminated, the learning curve would go back to an increasing state.

4 Raising the compatibility

Although both corpora include protein mention annotations, the target task is different. *GENIA concerns all the protein-mentioning terms, while AImed focuses only on the references of individual proteins*. In the GENIA corpus, besides the 36 classes, some subclasses are also included. In the case with the protein class, there are seven subclasses: family_or_group, domain_or_region, molecule, substructure, subunit, complex, etc. Further, in the AImed corpus, protein/gene families are not tagged, only protein molecules are tagged.

We conducted an experiment to verify what we found from the documentation of the two corpora. We trained our tagger using the AImed corpus, and evaluated it on the GENIA corpus. Each time, we assumed only the annotation of one protein subclass in the GENIA corpus as the "gold" annotation. Table 1 shows the experimental results.

The experimental results clearly supported the documented scope of the protein annotation in GENIA and AImed: The protein mention recognizer trained with AImed best recognized the GENIA annotation instances of Protein_molecules among all subclasses, and the performance of recognizing Protein_family_or_group instances was very poor.

We therefore have a hypothesis: if we unite the GENIA annotations of Protein_molecule, Protein_subunit, and Protein_complex with the AImed corpus, and we use this united corpus to train our tagger, we can improve the performance of our tagger on the AImed corpus. Table 2 shows our experimental results based on this hypothesis. It can be seen from the result that, if we assume that the upper bound of the F-score of this approach is near to 83.23%, we reduced the incompatibility of the two corpora by 30%. The reduction was obtained by understanding the difference of the protein annotations made to the corpora.

5 Conclusion

We implemented several experiments in order to remove the negative influence of the disagreements between two corpora. Our objective is to raise the compatibility of heterogeneous annotations. Some simple experiments partly revealed where the heterogeneity between the protein mention annotations in GENIA and AImed is. More qualitative and quantitative analysis will be done to identify the remaining heterogeneity.

References

Razvan Bunescu and Raymond Mooney. 2006. Subsequence Kernels for Relation Extraction. *Advances in Neural Information Processing Systems*, 18:171–178.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi and Jun'ichi Tsujii. 2003. GENIA Corpus - a Semantically Annotated Corpus for Bio-textmining. *Bioinformatics*, 19(Suppl. 1):i180–i182.