社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2008－ＮＬ－185　(8)
2008－SLP－ 71　(8)
2008／5／23

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

The use of human-annotated corpora is popular in developing language processing systems. For bio-text mining, for example, there are several well-known corpora with protein mention annotations. Because of the different conventions adopted by these corpora, one problem that is well recognized, but yet less addressed is brought about; the problem is the heterogeneity of the corpora. The problem weakens the protein annotation consistency. In this work, we seek a way of removing or relaxing the heterogeneity of annotations by identifying and removing the specific difference between the annotations. Our results show that our effort to remove several disagreements between the corpora annotation is successful in reducing the performance degradation caused by heterogeneity and incompatibility.

1. Introduction

The use of human-annotated corpora is widespread in developing language processing systems. For bio-text mining, for example, there are several well-known corpora with protein mention annotations: GENIA (Kim et al., 2003), PennBioIE (Mandel, 2006), GENETAG (Tanabe et al., 2005), AImed (Bunescu and Mooney, 2006), etc. Because of these corpora, many automatic protein mention recognizers have been developed, some of which report state-of-the-art performance (Ananiadou and McNaught, 2006; Jensen et al., 2006; Krallinger and Valencia, 2005; Wilbur et al., 2007; Yeh et al., 2005).

One of the remaining problems that is prominently recognized, but less studied, is regarding the compatibility of different annotations made to the corpora. The protein mention annotations to the above corpora were made by different groups with different conventions, resulting in heterogeneous and incompatible annotations, even though they are all supposed to represent the same task: identifying protein mentions from biomedical texts.

The heterogeneity of the annotations raises several problems. These include, but are not limited to:

- The performance of protein mention recognizers based on different annotations cannot be directly compared (Hersh, 2005).

- Although we have protein mention annotation made to enough sentences, we can utilize only a part of them at a time, and cannot mix heterogeneous annotations.

It is clear that by raising the compatibility of heterogeneous annotations, we can be much more efficient in developing expensive resources.

In this work, we seek a way of removing or relaxing the heterogeneity of annotations by identifying and removing the specific differences between them. We assumed a specific situation: one developed a protein mention recognizer based on one corpus, but soon found he/she wants more annotations to improve the performance. If he/she finds a way of utilizing annotations from another different corpus, he/she can save the enormous cost that it would take to perform additional manual annotations.

For exactly such a purpose, we explore the differences between two corpora, and design a series of experiments to see the effect of removing or relaxing the difference. Experimental results show that if we understand where the difference is, we can raise the compatibility of heterogeneous annotations by removing the difference.

2. Data

Here are several corpora mentioned in the previous section. Two of them are used in our work: GENIA corpus and AImed corpus. We will give a brief introduction on them focusing on the size and the annotation conventions of these two corpora.

2.1. GENIA corpus

GENIA corpus version 3.02 is a collection of articles extracted from the MEDLINE database with the MeSH terms, human, blood cell and transcription factor. There are 2,000 abstracts and 18,545 sentences altogether. Annotation is dependent on a small taxonomy of 48 classes based on a chemical classification. Among the classes, 36 terminal classes were used to annotate the GENIA corpus. The total number of recovered terms is 93,293. A simplified version called JNLPBA corpus (Kim et al., 2004) is also used in biotext mining domain.

2.2. AImed corpus

AImed corpus consists of 225 MEDLINE abstracts (1,969 sentences), and there are 4,084 protein references in this data set. Further, there is no distinction between genes and proteins.

3. Protein mention recognizer

Our protein mention recognition system is composed of an maximum entropy Markov model (MEMM) n-best tagger.

4. Preliminary experiments

This section describes a series of preliminary experiments carried out to characterize the problems that we are trying to deal with.

4.1. Experiment with AImed corpus

We implemented our first experiment using the AImed corpus for both training and evaluation. We divided the data into twenty contiguous and equally-sized sections. We then used the first fourteen sections for training, the following four sections for testing. In order to increase the size of the training corpus gradually, we made seven training subparts. Train 1 included the first two sections, and each time we added two following sections into the previous training subpart for the succeeding training, until we finished all seven training experiments. Table 1 shows the performance of our system on the AImed corpus according to exact match, left boundary match and right boundary match, respectively.

4.2. Experiment with the mixed corpus

In the previous experiment, we have suggested that we can improve the performance by increasing the size of the training data set. We are considering to concatenate the AImed corpus and the GENIA corpus, because of the sufficient size of the GENIA corpus. However, as we have mentioned earlier, the heterogeneity between two differently annotated corpora will absolutely cause noise. We cannot expect to solve such an annotation disagreement problem with some simple or manual rules.

(%)	Criterion	Recall	Precision	F-score
Train 1	Exact	54.41	69.27	60.94
	Left	60.41	76.91	67.67
	Right	59.13	75.28	66.24
Train 2	Exact	65.90	73.09	69.31
	Left	72.29	80.17	76.02
	Right	69.60	77.20	73.20
Train 3	Exact	69.22	75.28	72.12
	Left	74.97	81.53	78.11
	Right	72.67	79.03	75.72
Train 4	Exact	72.80	78.73	75.65
	Left	77.39	83.70	80.42
	Right	75.35	81.49	78.30
Train 5	Exact	72.67	81.05	76.63
	Left	77.39	86.32	81.62
	Right	75.10	83.76	79.19
Train 6	Exact	74.97	81.87	78.27
	Left	79.18	86.47	82.67
	Right	77.01	84.10	80.40
Train 7	Exact	76.12	82.09	78.99
	Left	80.20	86.50	83.23
	Right	78.16	84.30	81.11

Table 1: Recall, precision and F-score of the AImed_based experiment.

In order to demonstrate the effect of the incompatibility of the two corpora, we performed a simple experiment. In this experiment, we retrained our system by using a united corpus: the AImed corpus plus the GENIA corpus. We utilized all the protein subcategory annotations in the GENIA corpus[1], and treated all of these subcategories as positive examples in the training process. The training corpus from AImed is the same as the training corpus used in Train 7 from the previous experiment. The evaluation is done on the same part used for the evaluation in the earlier experiment.

Table 2 shows the recall, precision and F-score of the experimental result. We reason out that the performance is getting worse because of the introduced heterogeneity between two corpora.

Criterion	Recall	Precision	F-score
Exact	55.56	61.35	58.31
Left	64.88	71.65	68.10
Right	58.88	65.02	61.80

Table 2: Recall, precision and F-score of the experiment with the mixed corpus.

[1] We will explain it in details in the following section.

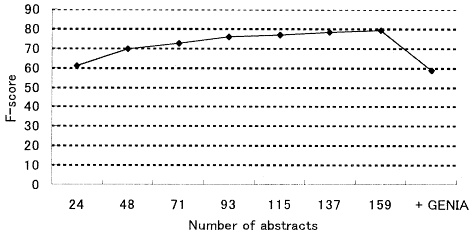

Figure 1: Learning curve according to the F-score of the exact match on outermost tags.

The learning curve drawn from both mentioned experimental results are shown in Figure 1. We can see from the learning curve that the learning curve is not yet saturated until we have used up all the training portions of the AImed corpus. The performance would be further improved as the size of the training data increases. But when we added the protein annotation in the GENIA corpus to the training material, we witnessed a drastic degradation in performance. We assume that the degradation is caused by the heterogeneity of the protein annotation in the two corpora, and we further assume that if the heterogeneity could be eliminated, the learning curve would go back to an increasing state.

5. Difference between GENIA and AImed

As previously mentioned, the incompatibility of heterogeneous corpora causes a serious problem when we try to use the corpora in an integrated way. Section 4. demonstrated the problem via a series of experiments. In this section, we explain some noticeable differences of the two corpora found from documentation and analysis.

Although both corpora include protein mention annotation, the target task is different. GENIA annotation centers on mining literature for general knowledge in biology, while AImed focuses on extracting interactions among individual proteins. The difference has affected the scope of annotated proteins: *GENIA concerns all the protein-mentioning terms while AImed focuses only on references of individual proteins.*

The scope of proteins annotated in the GENIA corpus is defined in the GENIA ontology (Ohta et al., 2002); besides protein class, other classes such as DNA, RNA, cell_line and cell_type are also included. Further the protein class is subcategorized into seven sub-classes: family_or_group, domain_or_region, molecule, substructure, subunit, complex, etc. In other words, in GENIA, the protein is defined as to include all the seven concepts.

In the case of AImed, the scope of the proteins annotated is described by the following statement in the AImed tagging conventions (Bunescu et al., 2005): generic protein/gene families are not tagged, only specific names that could ultimately be traced back to specific genes in the human genome are tagged.

6. Quantifying the difference

According to the differences between the two corpora, we implemented several sub-experiments so as to eliminate or relieve the degradation caused by these differences.

In section 5., we described the scope of the proteins of interest in the GENIA corpus and in the AImed corpus, respectively. Documentation of the two corpora explicitly states that:

(1) the mentions of protein families are annotated in GENIA, but not in AImed, and

(2) individual proteins (protein molecules) are annotated in both corpora.

These are the clues concerning the differences between the two corpora, which we were able to find from published documents. There are, however seven other subcategories of protein mentions annotated in the GENIA corpus, of which we could not find any mention regarding the inclusion or exclusion in the scope of proteins to be annotated in AImed. In GENIA, since the annotated protein mentions are subclassified into seven classes, we performed a series of experiments to confirm the two clues that we found from documents, and to find other clues for other protein subclasses. We trained our protein mention recognizer using only the AImed corpus, and applied it to a randomly selected part from the GENIA corpus, whose size is ten percent of GENIA, and compared the recognized results with the manual annotation. We evaluated the result of the recognition according to each subclass of protein at each time. In other words, at every time, we assumed only the annotation instances of one subclass of proteins as the "gold" annotation instances, and evaluated the performance of protein mention recognition in terms of how well it found the "gold" annotation instances. Table 3 shows the evaluated performance.

The evaluated performance based on each of Protein_family_or_group and Protein_molecule subclasses clearly supports the documented scope of the protein annotation in GENIA and AImed: The protein mention recognizer trained on the AImed corpus best recognized the GENIA annotation instances of Protein_molecules among all subclasses, and the per-

Subcategory	Answers / Instances	Recall	Precision	F-score
Family_or_group	114 / 881	12.94	3.86	5.94
Domain_or_region	17 / 108	15.74	0.57	1.11
Molecule	1018 / 2086	48.80	34.43	40.37
Substructure	0 / 17	0.00	0.00	0.00
Subunit	100 / 153	65.36	3.38	6.43
Complex	29 / 216	13.43	0.98	1.83
ETC	1 / 7	14.29	0.03	0.07
all	1279 / 3468	36.88	43.25	39.81

Table 3: Evaluated performance based on seven protein subcategories. In the last row, "all" means that we will think the identified entity as a TP if it is tagged as any one of seven subcategories in the GENIA corpus. In the second column, answers denote the number of correctly identified entities, and instances denote the number of annotated entities in the GENIA corpus.

formance of recognizing Protein_family_or_group instances was very poor. Also, the evaluated performance based on other sub-classes gives us some clues about other classes: protein mentions classified as Domain_or_region or ETC in GENIA might be out of scope of annotation in AImed, considering the low number of true positives (TPs), and so on.

In order to directly demonstrate the effect of using the GENIA annotation of each protein subclass, we performed another series of sub-experiments: we used the GENIA annotation of each subclass together with AImed for the training, and applied it to recognize protein mentions in the AImed corpus. Table 4 lists the results, which confirm our observations in Table 3. The GENIA annotation of the Protein_molecule most positively affected the performance of recognizing the proteins tagged in AImed corpus, and recognizing Protein_subunit and Protein_complex follows it. Note that the annotation of Protein_substructure and Protein_ETC were excluded from consideration since the number of corresponding examples is too small (17 and 7).

7. Raising compatibility

The experimental results discussed in the last section show that adding the GENIA protein annotation of Protein_molecule, Protein_subunit and Protein_complex separately to the training material improves the precision of the protein mention recognition on the AImed corpus at a significant cost of recall. This observation suggests that if we use all the three protein sub-classes as the training material, we could improve the recall while maintaining the level of precision. Table 5 shows our experiments on this hypothesis. It shows that when we collectively used the GENIA annotations of the three protein subclasses, the recall was improved significantly while minimizing the decrease in precision. Compared to the experimental result in Table 2, it is a significant improvement. When we assume that the upper bound of the F-score of this approach is near to 83.23% (left boundary matching), it can be said that we reduced the incompatibility of the two corpora by 30%. The reduction was obtained by understanding the difference of protein annotations made to the corpora.

8. Conclusion

We implemented several experiments in order to get rid of the bad influence of disagreements in annotation conventions. Our objective is to raise the compatibility of heterogeneous annotations. As we have already known, a system cannot perform well on another corpus when there are some distinctions in annotation rules between the training corpus and the testing corpus. And also as shown from our first experiment (section 4.1.), the performance will be improved by increasing the size of the training corpus. Thus, we believe that we can add another more comprehensive corpus (the GENIA corpus) into the previously used training corpus (the AImed corpus) to increase the performance. But the annotation difference must be eliminated in advance; otherwise, much noise will be generated and the performance will get worse rather than get better. To verify this, we implemented an experiment with a mixed training corpus (as shown in section 4.2.). All the protein subcategories are regarded as positive examples in the training. In this way, the experimental results show that the performance is getting worse, because of the introduced heterogeneity between two corpora. Further, we analyzed what kind of subcategories affects on the system performance the most by some subcategory based experiments (as shown in section 6. and in section 7.). We came to the conclusion that the incompatibility of heterogeneous

AImed + Subcategory	Matching criteria	Recall	Precision	F-score
Family_or_group	Exact	29.76	64.90	40.81
	Left	33.33	72.70	45.71
	Right	30.40	66.30	41.68
Domain_or_region	Exact	32.57	77.98	45.95
	Left	35.38	84.71	49.91
	Right	32.95	78.90	46.49
Molecule	Exact	52.75	82.60	64.38
	Left	55.81	87.40	68.12
	Right	55.17	86.40	67.34
Substructure	Exact	30.14	86.45	44.70
	Left	32.06	91.94	47.54
	Right	30.40	87.18	45.08
Subunit	Exact	33.72	80.49	47.52
	Left	36.91	88.11	52.03
	Right	34.23	81.71	48.24
Complex	Exact	34.99	80.12	48.71
	Left	39.21	89.77	54.58
	Right	35.89	82.16	49.96
ETC	Exact	33.84	87.75	48.85
	Left	35.50	92.05	51.24
	Right	34.10	88.41	49.22

Table 4: Recall, precision and F-score of sub-experiments based on subcategories.

AImed + Subcategory	Matching criteria	Recall	Precision	F-score
Molecule + Subunit	Exact	53.90	81.00	64.72
	Left	57.85	86.95	69.48
	Right	56.32	84.64	67.64
Molecule + Subunit + Complex	Exact	55.17	75.52	63.76
	Left	62.96	86.19	72.77
	Right	58.49	80.07	67.60

Table 5: Recall, precision and F-score of experiments on three protein subcategories.

annotations can be reduced by understanding where the difference is, and by properly considering the difference.

9. References

Sophia Ananiadou and John McNaught. 2006. Text Mining for Biology and Biomedicine. *Artech House*, London, UK.

Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun K. Ramani and Yuk Wah Wong. 2005. Comparative Experiments on Learning Information Extractors for Proteins and their Interactions. *Artificial Intelligence in Medicine*, 33:139–155.

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

Kevin Bretonnel Cohen, Lynne Fox, Philip V. Ogren and Lawrence Hunter. 2005. Corpus Design for Biomedical Natural Language Processing. *in Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases*, Detroit, USA.

William Hersh. 2005. Evaluation of Biomedical Text-mining Systems: Lessons Learned from Information Retrieval. *Briefings in Bioinformatics*, 6(4):344–356.

Lars Juhl Jensen, Jasmin Saric and Peer Bork. 2006. Literature Mining for the Biologist: from Information Retrieval to Biological Discovery. *Nature Publish Group*, 7:119–129.

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

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka and Yuka Tateisi. 2004. Introduction to the Bio-Entity Recognition Task at JNLPBA. *in Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, Geneva, Switzerland.

Martin Krallinger and Alfonso Valencia. 2005. Text-Mining and Information-Retrieval Services for Molecular Biology. *Genome Biology*, 6:224–231.

Mark A. Mandel. 2006. Integrated Annotation of Biomedical Text: Creating the PennBioIE Corpus. *in Proceedings of the Workshop on Text Mining, Ontologies and Natural Language Processing in Biomedicine*, Manchester, UK.

Tomoko Ohta, Yuka Tateisi, Hideki Mima and Jun'ichi Tsujii. 2002. GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain. *in Proceedings of the Human Language Technology Conference*, San Diego, USA.

Lorraine Tanabe, Natalie Xie, Lynne H Thom, Wayne Matten and W John Wilbur. 2005. GENETAG: a Tagged Corpus for Gene/Protein Named Entity Recognition. *BMC Bioinformatics*, 6(Suppl 1):S3–S9.

John Wilbur, Larry Smith and Lorrie Tanabe. 2007. BioCreative 2. Gene Mention Task. *in Proceedings of the Second BioCreative Challenge Evaluation Workshop*, Madrid, Spain.

Alexander Yeh, Alexander Morgan, Marc Colosimo and Lynette Hirschman. 2005. BioCreAtIvE Task 1A: Gene Mention Finding Evaluation. *BMC Bioinformatics*, 6(Suppl 1):S2–S11.