Exploring the Compatibility of Heterogeneous Protein Annotations Toward Corpus Integration

Yue Wang[*] **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, jdkim, rune.saetre, tsujii}@is.s.u-tokyo.ac.jp

Abstract

We explore the sources of incompatibility between the protein annotations made to two corpora: GENIA and AIMed. We first hypothesize a problem with the incompatibility caused by corpus integration, and we measure the effect of the incompatibility on protein mention recognition. Through a series of experiments, we find several sources of the incompatibility, and suggest that more than half of the incompatibilities can be reduced by properly considering the scope of the annotated proteins, text preprocessing, and boundary annotation conventions.

1 Introduction

Human-annotated corpora are widely used in developing language processing systems. For biotext mining, 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 et al., 2005), etc. Based on these corpora, many protein mention recognizers have been developed, some of which report state-of-the-art performance (Wilbur et al., 2007).

However, there remains a well-known, but less studied, problem. Since the protein annotations are made by different groups, it is likely that the annotations in different corpora are not compatible with each other.

The incompatibility brings about several significant problems. For example, it is difficult to effectively utilize more than one corpus to develop a protein mention recognizer. Indeed, there has never been a protein recognizer developed by utilizing multi-corpora, because it is hardly possible to benefit from corpus integration. It is also difficult to compare systems developed with different corpora. Although there are many systems that claim to recognize protein mentions from MEDLINE texts, their reported performance varies significantly (Tsai et al., 2006). The mentioned problems are largely caused by the incompatibility of different protein annotations, and can not be solved effectively without understanding the differences in the annotations (Pyysalo et al., 2008).

In this paper, we explore the potential sources of incompatibility between two well-known corpora with protein annotations: GENIA and AIMed. We first characterize the incompatibility resulting from using the two corpora as a single resource. Then, we carefully study the documentation of the two corpora in order to figure out the sources of incompatibility. Through a series of experiments, we explore the possible sources, while finding reasonable ways to avoid the problems caused by the incompatibility of protein annotations. Experimental results show that it is feasible to reduce the incompatibility of the heterogeneous annotations by properly considering the differences. Meanwhile, we can get a comprehensive understanding of the two corpora, and take advantage of the annotations in both corpora, while minimizing the negative effects caused by their inconsistency.

The paper is organized as follows. In section 2, the two corpora used for exploration, GENIA and AIMed, are described. Two preliminary experiments characterizing the problem of combining two incompatible corpora are reported in section 3. From section 4 to section 6, the corpora's differences are explored regarding three aspects: the scope of the entities of interest, text preprocessing, and the conventions for boundary decisions, respectively. We propose a way to reduce the corpus inconsistency for each aspect. Following a final experiment on the remaining inconsistencies in section 7, our research is concluded in section 8.

2 Data

Here, we briefly introduce the GENIA and the AIMed corpora, focusing on their size and covered domain.

2.1 The GENIA corpus

The GENIA corpus (version 3.02) is a collection of articles extracted from the MEDLINE database with the MeSH terms "human", "blood cells" and "transcription factors". There are 2,000 abstracts and 18,545 sentences totally. The term annotation is according to a taxonomy of 48 classes based on a chemical classification. Among the classes, 36 terminal classes were used to annotate the corpus. The total number of annotated terms is 93,293.

In recent years, the GENIA corpus has become one of the most frequently used corpora in biomedical named entity recognition (Bio-NER) task (Cohen and Hersh, 2005).

2.2 The AIMed corpus

The AIMed corpus consists of 225 MEDLINE abstracts, of which 180 are known to describe interactions between human proteins, while the other 45 do not refer to any interaction. In all, there are 1,969 sentences and 4,084 protein references.

The AIMed corpus is now one of the most widely used corpora with protein interaction annotation. Its protein annotations are parts of the protein interaction annotations.

3 Preliminary experiments

We performed two preliminary experiments in order to confirm the following two assumptions. First, we can improve the performance of a protein mention recognizer by increasing the size of the training data set. Second, the system performance will drop when incompatible annotations are introduced into the training data set. The protein mention recognizer used in our work is a Maximum Entropy Markov Model n-best tagger (Yoshida and Tsujii, 2007). To reduce our task to a simple linear sequential analysis problem, we removed all the embedded tags in GENIA and AIMed, and only retained the outermost tags.[1]

We divided the AIMed corpus into two parts, 70% for training and the remainder for testing. In the first experiment, we only used the AIMed training part. In this experiment, we performed seven sub-experiments, and each time, we added 10% more abstracts into the training portion. In the second experiment, besides the AIMed training part, we also added the GENIA protein annotations. In both experiments, we performed the evaluations on the AIMed test part according to the exact matching criterion. In this paper, all the evaluations are carried on the AIMed test part, whose size is 30% of the AIMed corpus. For convenience, the AIMed training part is simply called the "AIMed corpus" in the following.

A learning curve drawn from the results of the two mentioned experiments is shown in Figure 1.

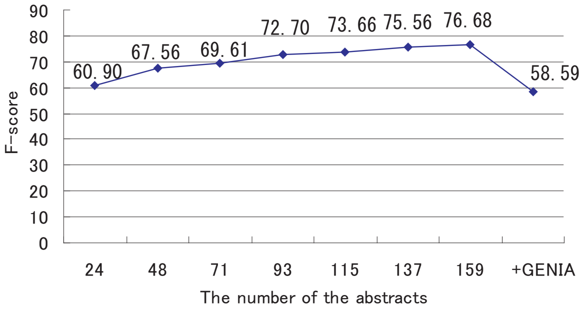

Figure 1: Learning curve drawn from the results of two preliminary experiments.

We can see that the learning curve is still increasing when we used up all the training portions of the AIMed corpus. We would expect a further improvement if we could add more training data in a large scale, e.g. the GENIA corpus, which is ten times bigger than the AIMed corpus. But when we actually add the protein annotations in the GENIA corpus to the training data set, we witness a drastic degradation in the performance. We assume that the degradation is caused by the incompatibility of the protein annotations in the two corpora, and we further assume that as the incompatibility decreases, the learning curve would get back to the original increasing direction.

In the following three sections, we will ex-

[1] There are 136 embedded occurrences in the AIMed corpus, 3 of which are triple-nested. And there are 1,494 embedded cases in the GENIA corpus.

plain some differences that cause the performance degradation, both from the perspective of documentation and from the experimental results. According to the differences, we design a series of experiments to reduce the incompatibility between the two corpora when using them in an integrated way.

4 Scope of the entities of interest

Although both corpora include protein mention annotations, the target tasks are different. The GENIA annotation centers on mining literature for general knowledge in biology, while the AIMed annotation 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.*

4.1 Categories of annotated proteins

The scope of the proteins annotated in the GENIA corpus is defined in the GENIA ontology (Ohta et al., 2002); besides the protein class, other classes such as *DNA*, *RNA*, *cell_line* and *cell_type* are also included. Further, the protein class is categorized into seven subclasses: *Protein_complex*, *Protein_domain_or_region*, *Protein_family_or_group*, *Protein_molecule*, *Protein_substructure*, *Protein_subunit* and *Protein_ETC*. In other words, in GENIA, the protein is defined to include all seven concepts. No other protein subclasses are defined in the GENIA corpus.

In the case of AIMed, the scope of the proteins annotated is described by the following statement in the tagging conventions: generic protein families are not tagged, only specific names (protein molecules) that could ultimately be traced back to specific genes in the human genome are tagged. E.g. “Tumor necrosis factor” would not be tagged, while “tumor necrosis factor alpha” would be.

Hence, the documentation of the two corpora explicitly states that:

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

(2) individual proteins (*Protein_molecule*) are annotated in both corpora.

4.2 Compatible annotations

Section 4.1 provided two clues for the inclusion/exclusion of *Protein_molecule* and *Protein_family_or_group* annotations, specified in the published literature. However, there are five other protein subcategories annotated in GENIA, and we could not find any mentions regarding the inclusion or exclusion of the five protein subcategories in the scope of the annotations in AIMed. We performed a series of experiments to confirm the two clues that we found, and to find other clues for the other five protein subclasses.

+ Subcategory	Recall	Precision	F-score
molecule	52.87	82.80	64.54
subunit	29.63	86.57	44.15
ETC	28.61	89.60	43.37
substructure	28.10	88.00	42.59
complex	28.48	79.93	42.00
domain_or_region	27.71	79.49	41.10
family_or_group	26.82	65.02	37.97

Table 1: Experimental results of the AIMed corpus plus the GENIA protein subcategory annotations.

We used each of the GENIA protein subclasses in turn together with the AIMed corpus for the training. That is, each time we regarded the annotations from a different GENIA protein subclass as positive examples. The experimental results are listed in Table 1, showing the exact matching scores. According to the table, it is most harmful to add the *Protein_family_or_group* annotations, supporting the clue we have already found: the mentions of protein families are annotated in GENIA, but not in AIMed. Also, we notice that the GENIA *Protein_molecule* annotations least negatively affect the performance of recognizing the proteins tagged in the AIMed corpus, and the *Protein_subunit* and *Protein_complex* follow it[2]. Meanwhile, we observe that by adding the protein subcategory annotations, the precision of the protein mention recognition on the AIMed corpus is very good, while the recall is very low. This observation suggests that if we add the annotations of the three protein sub-classes into the training material at the same time, we could improve the recall while maintaining good precision. Table 2 shows the experimental results based on this

[2] Because the number of the *Protein_substructure* annotations and the *Protein_ETC* annotations are very small (103 and 85, respectively), the two protein subcategories were excluded from consideration.

AIMed + Subcategory	Criterion	Recall	Precision	F-score
molecule + subunit	Exact	53.77	80.96	64.62
	Left	58.75	88.46	70.61
	Right	56.70	85.38	68.15
	Overlap	62.20	93.65	74.75
molecule + subunit + complex	Exact	54.15	76.40	63.38
	Left	62.58	88.29	73.24
	Right	57.34	80.90	67.12
	Overlap	67.05	94.59	78.47

Table 2: Experimental results of the AIMed corpus plus the annotations of three GENIA protein subcategories.

Training data	Criterion	Recall	Precision	F-score
AIMed	Exact	74.33	79.18	76.68
	Left	78.93	84.08	81.42
	Right	76.63	81.63	79.05
	Overlap	81.48	86.80	84.06
AIMed + GENIA_Protein	Exact	56.19	61.20	58.59
	Left	66.79	72.74	69.64
	Right	59.90	65.23	62.45
	Overlap	72.80	79.28	75.54

Table 3: Experimental results of the AIMed corpus and the AIMed corpus plus the GENIA protein annotations.

hypothesis. In addition to the exact, left boundary and right boundary matching criteria, we also tested an overlap matching criterion (Franzén et al., 2002), namely, if any part of a protein mention is identified, it will be considered as a correct answer. The experimental results show that when we collectively use the GENIA annotations of the three protein subclasses, the recall improved significantly while minimizing decrease in precision. For fair comparison, we also applied the left boundary, right boundary and overlap matching criteria to the results gained by using the AIMed corpus, and the AIMed corpus plus the GENIA protein annotations, respectively. The results are shown in Table 3.

Since our goal is to find a way to make the learning curve go back to an increasing state, we set the performance induced from the pure AIMed corpus as the minimum goal. Then, the potential (maximum) reduction rate of incompatibility can be calculated by Formula (1):

$$R_e = \frac{F_e - F_{A+G}}{F_A - F_{A+G}}\%, \qquad (1)$$

where R_e denotes the corpus incompatibility reduction rate of a given experiment, F_e denotes the F-score of the given experiment, F_A and F_{A+G} denote the F-score of the training with the AIMed corpus, and with the AIMed corpus plus the GENIA protein annotations, respectively.

We can say that, by combining the GENIA *Protein_molecule*, *Protein_subunit* and *Protein_complex* annotations with the AIMed corpus, we reduced the corpus incompatibility by 30.56% (the left boundary matching criterion[3]). So, when we want to introduce the annotations from the GENIA corpus, we can use the annotations of the three protein subclasses. It further indicates that the annotations of these three protein subclasses in both corpora are compatible to some extent.

We found sentences including *Protein_subunit* or *Protein_complex* annotations, which will not cause the incompatibility during corpus combination[4]. That is, in both corpora, these entities are regarded as proteins, so we can introduce most of the GENIA annotations of these entities into AIMed, without negative influence. Some examples are shown in Figure 2. For comparison, all the entity annotations are shown in the figure.

4.3 Ambiguity between DNAs and genes

The protein annotations in the AIMed corpus include not only proteins, but also genes, without differentiating them. In the case of the GENIA corpus, the protein annotation is applied

[3] To avoid underestimation, we adopted a looser criterion.

[4] *Protein_molecule* has already been annotated in both corpora.

Disruption of the $\dfrac{\textbf{Jak1 binding proline-rich Box1 region}}{\textit{Protein\_domain\_or\_region}}$ of $\dfrac{\textbf{IL-4Ralpha}}{\textit{Protein\_subunit}}$ abolished signaling by this $\dfrac{\textbf{chimeric receptor}}{\textit{Protein\_family\_or\_group}}$. (GENIA PMID 9159166)

Only weak **IL-13** binding activity was found in cells transfected with only **IL-13Ralpha**; however, the combination of both **IL-13Ralpha** and **IL-4Ralpha** resulted in substantial binding activity, with a Kd of approximately 400 pM, indicating that both chains are essential components of the **IL-13 receptor**. (AIMed PMID 8910586)

Triflusal and HTB may exert beneficial effects in processes in which de novo $\dfrac{\textbf{COX-2}}{\textit{Protein\_molecule}}$ expression is involved and, in a broader sense, in pathological situations in which genes under $\dfrac{\textbf{nuclear factor-kappaB}}{\textit{Protein\_complex}}$ control are up-regulated. (GENIA PMID 10101034)

In this review, we summarize these and other TNF receptor-associated proteins and their potential roles in regulating the activation of **nuclear factor-kappaB** and apoptosis, two major responses activated by engagement of TNF receptors by the ligand. (AIMed PMID 9129204)

Figure 2: Sentences including the same annotated entities. (The boldface represents an annotated entity and in the GENIA examples the word under the line represents the class used to annotate the entity.)

only to proteins, while genes are annotated in the scope of DNA annotations. This suggests that it would improve the consistency if we treat gene annotations in the GENIA corpus in the same way as done in the AIMed corpus. However, the GENIA annotation does not include an explicit gene annotation. Instead, genes are annotated as instances of *DNA_domain_or_region* which is also applied to other DNA regions; e.g. binding sites and c-terminals. We assume that if the *DNA_domain_or_region* annotations that are not pure genes can be filtered out from all the *DNA_domain_or_region* annotations, we can find some examples from the remaining GENIA *DNA_domain_or_region* annotations that will positively affect the corpus combination. Therefore, if we assume that the performance of the recognizer trained with the AIMed corpus is good enough,[5] it will find most of the gene mentions in the GENIA corpus. The true positives, which are annotated as *DNA_domain_or_region* in the GENIA corpus and are also recognized by the recognizer, will include *DNA_domain_or_region* instances which are genes.

To examine the performance of the filtering, we added all the *DNA_domain_or_region* annotations to the training set in one experiment, and only the "true positive" classified "genes" in another experiment. The results shown in Table 4 indicate that the disambiguation between DNAs and genes works, although the improvement degree resulting from the filtering is not big.

As mentioned in section 4.2, adding only the *Protein_molecule*, *Protein_subunit* and *Protein_complex* annotations gives the best performance on the AIMed test part. Next, besides these three annotation types, we also added the filtered *DNA_domain_or_region* annotations to train our protein mention recognizer. The experimental results are shown in Table 5. Compared with Table 3, the corpus incompatibility is reduced 40.58% by adding the filtered D*NA_domain_or_region* annotations (the left boundary matching criterion).

AIMed + Subcategory	Recall	Precision	F-score
DNA	29.76	80.62	43.47
DNA_which_is_a_gene	30.27	84.95	44.63

Table 4: Experimental results of the disambiguation between *DNA_domain_or_region* and gene based on the exact matching criterion.

Criterion	Recall	Precision	F-score
Exact	56.58	74.70	64.39
Left	65.39	86.34	74.42
Right	60.28	79.60	68.60
Overlap	70.63	93.25	80.38

Table 5: Experimental results of adding the *Protein_molecule*, *Protein_subunit* and *Protein_complex*, and the filtered *DNA_domain_or_region* annotations.

[5] Of course, the filtering would only work perfectly, on the premise that the performance of the recognizer is perfect, so it will be a rough filtering.

5 Text preprocessing

In the AIMed corpus, a pre-tokenization policy is taken, which is the Penn Tree Bank style tokenization. Hence, we also pre-tokenized the GENIA corpus according to the Penn Tree Bank style, and retrained our recognizer by combining the *Protein_molecule_subunit_complex* annotations, and the filtered *DNA_domain_or_region* annotations, with the AIMed corpus. The experimental results are shown in Table 6. Compared with the results from Table 3, we reduce the incompatibility of the two corpora by 44.57% (the left boundary matching criterion).

Criterion	Recall	Precision	F-score
Exact	58.75	75.29	66.00
Left	66.67	85.43	74.89
Right	61.56	78.89	69.15
Overlap	70.88	90.83	79.62

Table 6: Experimental results of taking Penn Tree Bank style pre-tokenization.

6 Boundary of protein mentions

Even though the scope of the proteins to be annotated is standardized, the boundary of the protein mentions is still ambiguous. In general, the boundary ambiguity often arises in two ways. One ambiguity exists in making general guidelines for which part of a text expression is in charge of mentioning a protein. The other ambiguity exists regarding the confusion concerning the application of these guidelines. The confusion can be measured by entropy, as described below.

6.1 Determining which part is in charge of protein mentions

For example, when the text expression "p21ras protein" is given, it is not obvious whether to annotate the word "protein" as a part of the protein mentioning expression or not. We found that GENIA includes the word "protein" in the protein mentioning expressions, while AIMed excludes it. If a Bio-NER system is trained with AIMed, and we evaluate this system on GENIA, we can see a boundary matching error in "p21ras" , where "protein" is not included in the tag. However, for a text mining system, this error may be acceptable, since the system has correctly identified "p21ras" as a protein, and this information is adequate to mine the relationship between "p21ras" and another protein. Similarly, "the p21ras protein" or "the p21ras" could also be considered correct.

This also affects the average length of the protein mentions in the two corpora. The average length per protein mention is 1.9 tokens in the AIMed corpus, and 2.9 in the GENIA corpus. The percentages of protein mentions over 3 tokens in AIMed and GENIA are 12.65% and 50.29%, respectively. Many long protein mentions are introduced when we add the GENIA annotations into AIMed; this is another source of the performance degradation of recognizing shorter protein mentions in the AIMed corpus.

6.2 Annotation entropy for boundary words

In a given corpus, some words are annotated as inside of protein mentions, while other words are not. The annotation entropy of boundary words is calculated by Formula (2). For the sake of brevity, the (boundary) "word" discussed in this subsection describes the word that appears at the beginning or end of an annotated entity, or that abuts an annotated entity. When the annotation entropy of a boundary word is 0, this word is perfectly annotated and keeps the annotation consistency in the entire corpus. On the contrary, when the annotation entropy of a boundary word is 1, this word is so disorderly annotated that we can hardly find any rules about whether to regard it as a part of protein mentions or not. The value of E_b ranges from 0 (consistent) to 1 (inconsistent).

$$E_b = -(P_a \log_2 P_a + \overline{P_a} \log_2 \overline{P_a}), \qquad (2)$$

where, E_b denotes the annotation entropy of a given word, P_a denotes the percent of the annotated occurrences of this word, and $\overline{P_a}$ denotes the percent of the occurrences of this word that are not annotated.

In general, there are two types of boundary words: descriptive adjectives (such as "normal" or "activated"), and nouns, denoting the semantic category, occurring either before (as modifiers, such as "human") or after (as heads, such as "protein" or "molecule") The GENIA tagger[6] was used to determine the words Part-Of-Speech. Some boundary words appearing in each corpus are listed in Table 7. In order to characterize the differences between the two corpora in terms of

[6] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

Category	Word	AIMed			GENIA		
		N_a	N_n	E_b	N_a	N_n	E_b
Adjective	recombinant	1	7	0.54	36	24	0.97
	soluble	1	9	0.48	14	15	1.00
	inducible	0	0	0.00	18	18	1.00
	putative	0	0	0.00	15	15	1.00
	constitutive	0	0	0.00	12	11	1.00
	low	0	0	0.00	14	11	0.99
	major	0	0	0.00	25	15	0.95
Noun_before	protein	12	29	0.73	164	18	0.47
Noun_after	protein	36	17	0.96	749	45	0.31
	site	0	0	0.00	21	12	0.95

Table 7: List of boundary words. Here, Noun_before indicates the noun occurring before an entity as a modifier, Noun_after indicates the noun occurring after an entity as a head. N_a is the number of the annotated occurrences, and N_n is the number of not annotated occurrences.

annotation entropy of boundary words, the words with annotation entropy close to 1 in one of the two corpora were included in Table 7.

From the table, we can see that the boundary annotation problem appears for various words. The distribution of these words is diverse, especially for the case of adjectives. Since as few extra characters as possible were tagged in the AIMed corpus, only the names of protein mentions are annotated, and most of the adjectives are not annotated. However, in the GENIA corpus, the adjectives before protein mentions are annotated only if they are required for the meaning of protein mentions (e.g. in the protein mention of "inducible cAMP early repressor", "inducible" is annotated, because it is needed for the comprehension of the meaning.).

In this situation, we need an alternative matching criterion other than the exact matching. To provide alternative evaluation perspectives, researchers have developed a variety of evaluation criteria that relax the matching to different degrees. Here, as previously shown, in addition to exact matching, left boundary, right boundary, and overlap matching are considered. Thus, if we assume that the expected minimal performance of the F-score of this work is near 84.06% (no corpus integration), it can be said that the possible maximum reduction of the incompatibility between the two corpora by the methods in this paper is 56.81% (the overlap matching criterion in Table 5).

7 Experiments performed on the non-overlapped data

From our current best results shown in section 6, there are still remaining incompatibilities responsible for more than half of the total incompatibilities. Since the abstracts in the two corpora are collected in different ways, it is supposed that the proteins mainly mentioned in the two corpora are heterogeneous, resulting in the incompatibility.

To quantify this assumption, we counted the number of identical names between the training and the testing part of AIMed, and between the AIMed training part and the GENIA protein annotations. In order to find the unique proteins, we normalized the protein mentions; for example, we changed uppercases to lowercases, and removed punctuation marks, spaces, and the appositions in parentheses. There are 766 unique entities in the AIMed training part, 250 in the AIMed test part and 7,759 in the GENIA corpus. Between the AIMed training part and the GENIA protein annotations, there are merely 270 unique entities that are overlapped. Further, between the training and the test parts of AIMed, the number of the overlapped unique entities is just 91. Due to the low overlapping coefficient, we divided the AIMed test part into two parts: one includes the annotations overlapped with the AIMed training part, and another includes the annotations that are not overlapped with the AIMed training part. We re-evaluated our recognizer on the latter part. The experimental results are shown in Table 8. From the table, even though the performance on the non-overlapped part did not improve by adding the three GENIA protein subcategories and the filtered *DNA_domain_or_region* annotations, and by taking pre-tokenization, the result is very close to the result gained by using only the AIMed corpus for training. It implies that the heterogeneity of the proteins in the two corpora is another major source of the incompatibility, and suggests that we need find an appropriate way to properly consider the heterogeneity.

Training data	Recall	Precision	F-score
AIMed	71.46	40.54	51.73
AIMed+GENIA	63.31	43.21	51.36

Table 8: Experimental results of the non-overlapped part based on the overlap matching criterion. The last row shows the result of adding the three GENIA protein subcategories and the filtered *DNA_domain_or_region* annotations, and taking pre-tokenization.

8 Conclusions

Incompatibility of protein annotations in different corpora is a well known, but less studied, problem. In order to measure the effect of the incompatibility on protein mention recognition, we performed an experiment of corpus integration, which showed a significant degradation of performance due to the incompatibility.

Motivated by the result of the preliminary experiment, we investigated the source of incompatibility through a series of experiments. The results were encouraging. We found three main sources of incompatibility: the scope of the entities of interest, text preprocessing, and boundary of protein mentions, thus suggesting ways of reducing or avoiding the incompatibility. Meanwhile, we could improve our understanding of the difference of the two corpora, leading to a better understanding about the performance of protein recognizers based on them.

Some future works will follow from two perspectives. In order to achieve an actual improvement of protein recognition by integrating different corpora, we will further investigate the remaining source of incompatibility, finding a suitable model to integrate heterogeneous annotations. In order to better understand the difference of protein annotations, we will extend the comparison work to other corpora, e.g. GENETAG, toward a better consensus of protein annotations.

Acknowledgments

This work was partially supported by Grant-in-Aid for Specially Promoted Research (MEXT, Japan) and Genome Network Project (MEXT, Japan).

References

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.

Aaron M. Cohen and William R. Hersh. 2005. A Survey of Current Work in Biomedical Text Mining. *Briefings in Bioinformatics*, 6:57–71.

Kristofer Franzén, Gunnar Eriksson, Fredrik Olsson, Lars Asker, Per Lidén and Joakim Cöster. 2002. Protein Names and How to Find Them. *International Journal of Medical Informatics*, 67:49–61.

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.

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.

Sampo Pyysalo, Antti Airola, Juho Heimonen, Jari Björne, Filip Ginter and Tapio Salakoski. 2008. Comparative Analysis of Five Protein-protein Interaction Corpora. *BMC Bioinformatics*, 9(Suppl 3):S6–S16.

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(S1):S3–S9.

Richard Tzong-Han Tsai, Shih-Hung Wu, Wen-Chi Chou, Yu-Chun Lin, Ding He, Jieh Hsiang, Ting-Yi Sung and Wen-Lian Hsu. 2006. Various Criteria in the Evaluation of Biomedical Named Entity Recognition. *BMC Bioinformatics*, 7:92–99.

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

Kazuhiro Yoshida and Jun'ichi Tsujii. 2007. Reranking for Biomedical Named-Entity Recognition. *in Proceedings of the workshop of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic.