Incorporating GENETAG-style annotation to GENIA corpus

Tomoko Ohta[*] and **Jin-Dong Kim**[*] and **Sampo Pyysalo**[*] and **Yue Wang**[*] and **Jun'ichi Tsujii**[*†‡]
[*]Department of Computer Science, University of Tokyo, Tokyo, Japan
[†]School of Computer Science, University of Manchester, Manchester, UK
[‡]National Centre for Text Mining, University of Manchester, Manchester, UK
{okap,jdkim,smp,wangyue,tsujii}@is.s.u-tokyo.ac.jp

1 Introduction

Proteins and genes are the most important entities in molecular biology, and their automated recognition in text is the most widely studied task in biomedical information extraction (IE). Several corpora containing annotation for these entities have been introduced, GENIA (Kim et al., 2003; Kim et al., 2008) and GENETAG (Tanabe et al., 2005) being the most prominent and widely applied. While both aim to address protein/gene annotation, their annotation principles differ notably. One key difference is that GENETAG annotates the conceptual entity, gene, which is often associated with a function, while GENIA concentrates on the physical forms of gene, i.e. protein, DNA and RNA. The difference has caused serious problems relating to the compatibility and comparability of the annotations. In this work, we present an extension of GENIA annotation which integrates GENETAG-style gene annotation. The new version of the GENIA corpus is the first to bring together these two types of entity annotation.

2 GGP Annotation

Gene is the basic unit of heredity, which is encoded in the coding region of *DNA*. Its physical manifestations as *RNA* and Protein are often called its products. In our view of these four entity types, gene is taken as an abstract entity whereas protein, DNA and RNA are physical entities. While the three physical entity types are disjoint, the abstract concept, gene, is defined from a different perspective and is realized in, not disjoint from, the physical entity types.

The latest public version of GENIA corpus (hereafter "old corpus") contains annotations for gene-related entities, but they are classified into only physical entity types: Protein, DNA and RNA. The corpus revisions described in this work are two-fold. First, annotation for the abstract entity, gene, were added (Table 1, GGP). To emphasize the characteristics of the new entity type, which does not distinguish a gene and its products, we call it GGP (gene or gene product). Second, the addition of GGP annotation triggered large-scale removal of Protein, DNA and RNA annotation instances for cases where the physical form of the gene was not referred to (Due to space limitations, we omit RNA from now on). The time cost involved with this revision was approximately 500 person-hours.

	Protein	DNA	RNA	GGP
Old Annotation	21,489	8,653	876	N/A
New Annotation	15,452	7,872	863	12,272

Table 1: Statistics on annotation for gene-related entities

3 Quality Assessment

To measure the effect of revision, we performed NER experiments with old and new annotation (Tables 2 and 3). We split the corpus into disjoint 90% and 10% parts for use in training and test, respectively. We used the BANNER (Leaman and Gonzalez, 2008) NE tagger and created a separate single-class NER problem for each entity type.

In the old annotation, consistency is moderate for protein (77.70%), while DNA is problematic (58.03%). The new GGP annotation has been achieved in a fairly consistent way (81.44%). However, the removal of annotation for entities previously marked as protein or DNA had opposite effects on the two: better performance for DNA (64.06%),

Proceedings of the Workshop on BioNLP, pages 106–107,
Boulder, Colorado, June 2009.

	Precision	Recall	F-score
Protein	80.78	74.84	77.70
DNA	64.90	52.48	58.03

Table 2: NER performance before GGP annotation

	Precision	Recall	F-score
Protein	71.20	56.61	63.08
DNA	69.59	59.35	64.06
GGP	86.86	76.65	81.44
Protein+	83.22	78.20	80.63

Table 3: NER performance after GGP annotation

Phosphorylation		Gene_expression	
GGP_in_protein	70%	GGP_abstract	34%
Protein	25%	Protein	24%
GGP_abstract	3%	GGP_in_Protein	17%
Peptide	1%	GGP_in_DNA	9%

Table 4: Distribution of theme entity types in GENIA

implying annotation consistency improved with the removals, but worse for Protein (63.08%).

We find the primary explanation for this effect in the statistics in Table 1: in the revision, a large number of protein annotations (6,037) but only a small number of DNA annotations (780) were replaced with GGP. To distinguish such GGPs from those embedded in Protein or DNA annotations, we call them "abstract" GGPs, as they appear in text without information on their physical form. Nevertheless, in the old annotation, they had to be annotated as either protein or DNA, which might have caused inconsistent annotation. However, the statistics show a clear preference for choosing Protein over DNA. The radical drop of performance in protein recognition can then be explained in part as a result of removing this systematic preference.

Aside from the discussion on whether the preference is general or specific, we interpret the preference as a need for "potential" proteins to be retrieved together with "real" proteins, which was answered by the old protein annotation. To reproduce this class in the new annotation, we added abstract GGPs to the Protein annotation and performed an NER experiment. The result (Table 3, Protein+) shows a clear improvement over the comparable result for the old protein annotation.

In conclusion, we argue, the revision of the GENIA annotation, in addition to introducing a new entity class, has led to a significant improvement of overall consistency.

4 Discussion

Although there are already corpora such as GENETAG with annotation similar to GGPs, we expect this newly introduced class of annotation to support existing annotations of GENIA, such as event and co-reference annotation, opening up new possibilities for application. The quality of entity annotation should be closely related to that of other semantic annotation, e.g. events. For example, the event type Phosphorylation is about a change on physical entities, e.g. proteins and peptides, and as such, it is expected that themes of these events would be physical entities. On the other hand, the event type Gene_expression is about the manifestation of an abstract entity (gene) as a physical entity (protein) and would thus be expected to involve both abstract and physical entities. Statistics from GENIA (Table 4) show that the theme selection made in event annotation well reflects these characteristics of the two event types. The observation suggests that there is a good likelihood that improvement of the entity annotation can be further transferred to other semantic annotation, which is open for future work.

Acknowledgments

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

References

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.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from lterature. *BMC Bioinformatics*, 9(1):10.

R. Leaman and G. Gonzalez. 2008. Banner: an executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing*, pages 652–663.

Lorraine Tanabe, Natalie Xie, Lynne Thom, Wayne Matten, and W John Wilbur. 2005. Genetag: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6(Suppl 1):S3.