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.

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%),

106

Proceedings of the Workshop on BioNLP, pages 106–107, Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics
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 num-
ber of protein annotations (6,037) but only a small
number of DNA annotations (780) were replaced
with GGP. To distinguish such GGPs from those em-
bbeded in Protein or DNA annotations, we call them
“abstract” GGPs, as they appear in text without in-
formation on their physical form. Nevertheless, in
the old annotation, they had to be annotated as either
protein or DNA, which might have caused inconsis-
tent annotation. However, the statistics show a clear
preference for choosing Protein over DNA. The rad-
ical 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 pref-
ereence is general or specific, we interpret the pref-
ereence as a need for “potential” proteins to be re-
trieved together with “real” proteins, which was an-
swered 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 re-
sult for the old protein annotation.

In conclusion, we argue, the revision of the GE-
NIA annotation, in addition to introducing a new en-
tity class, has led to a significant improvement of
overall consistency.

4 Discussion

Although there are already corpora such as GENE-
TAG with annotation similar to GGPs, we expect
this newly introduced class of annotation to support
existing annotations of GENIA, such as event and
coreference annotation, opening up new possibili-
ties for application. The quality of entity annota-
tion should be closely related to that of other seman-
tic annotation, e.g. events. For example, the event
type Phosphorylation is about a change on physi-
ical 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 ab-
stract 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 anno-
tation well reflects these characteristics of the two
event types. The observation suggests that there is a
good likelihood that improvement of the entity an-
notation 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 an-
notated 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
literature. BMC Bioinformatics, 9(1):10.
R. Leaman and G. Gonzalez. 2008. Banner: an exe-
cutable survey of advances in biomedical named en-
tity recognition. Pacific Symposium on Biocomputing,
pages 652–663.
Lorraine Tanabe, Natalie Xie, Lynne Thom, Wayne Mat-
ten, and W John Wilbur. 2005. Genetag: a tagged cor-
pus for gene/protein named entity recognition. BMC
Bioinformatics, 6(Suppl 1):S3.