The Impact of Annotation on the Performance of Protein Tagging in Biomedical Text

Beatrice Alex, Malvina Nissim, Claire Grover

School of Informatics
University of Edinburgh
2 Buccleuch Place
EH8 9LW
Edinburgh, UK
{balex, mnissim, grover}@inf.ed.ac.uk

Abstract

In this paper we discuss five different corpora annotated for protein names. We present several within- and cross-dataset protein tagging experiments showing that different annotation schemes severely affect the portability of statistical protein taggers. By means of a detailed error analysis we identify crucial annotation issues that future annotation projects should take into careful consideration.

1. Introduction

The huge and constantly increasing amount of electronically available papers in the biomedical domain has triggered a high volume of research on automatically extracting biomedical information from the literature. A primary step in information extraction is recognising entities of interest, such as proteins, genes, cell lines, and so on, for which relevant information can then be extracted. Machine learning approaches currently dominate the field, and in order to train statistical models several text collections have been annotated with these entities. The attractiveness of learning methods is their portability and reusability: given the same entity types and similar text types, one should, for example, be able to use a protein name model trained on one corpus to recognise proteins in a new one.

However, different annotation projects employ different definitions of what constitutes an entity. As a consequence, what is marked as a protein in one text collection might not be marked as one in another, or their boundaries might not coincide. In this paper, we present a survey of different corpora and annotation schemes for protein markup and show that they differ substantially. The biomedical corpora which we examine are described in Section 2. By means of several within- and cross-dataset protein tagging experiments on such corpora, described in Section 3, we show that this lack of annotation standards constitutes a major shortcoming in the portability and reusability of trained models and their evaluation. We also present examples of different error types. In Section 4, we provide an overview of within- and cross-dataset inconsistencies in the annotation.

2. Biomedical Datasets and Preprocessing

With more and more biomedical datasets becoming publicly available, there has been some research effort on corpus design issues and usage in biomedical natural language processing (Cohen et al., 2005a; Cohen et al., 2005b). It is suggested that the usability and usefulness of a corpus is largely dependent on the annotation format and the provision of high quality markup of structural and linguistic characteristics of its content. In this paper, we look at the actual entity annotation variation within and across different collections and argue for standardised annotation guidelines for biomedical named entities.

We discuss experiments involving five different biomedical datasets, namely the BioNLP Coling 2004 corpus (Collier et al., 2004), the Texas UTML corpus (Bunescu et al., 2005), the PIR Georgetown corpus (Mani et al., 2005), the Yapex corpus (Franzén et al., 2002) and the Bio1 corpus (Tateisi et al., 2000). All five datasets are composed of (sentences from) Medline abstracts and contain manually annotated protein names.

We also examined the BioCreative data (Hirschman et al., 2004) and the oncology data of PennBioIE version 0.9 (Bies et al., 2005). The BioCreative corpus is made up of sentences selected from Medline abstracts. 50% of them were taken from abstracts similar to documents with known gene or protein names. The other 50% stem from abstracts unlikely to contain such names. In this dataset, protein names are not differentiated from gene names and are assigned the same entity tag. The PennBioIE oncology dataset contains Medline abstracts on the molecular genetics of cancer. This dataset differentiates between the entities Gene-RNA (genes and RNA elements), Gene-protein (non-genomic downstream products of genes and RNA elements) and Gene-generic (if unclear or ambiguous). Because we are interested in protein tagging as distinguished from gene tagging, we omit these two corpora from the experiments and discussion presented in this paper.

As the original markup of all datasets differs, we converted it into a common standard XML representation. We then tokenised all datasets with the same rule-based grammar which splits text into tokens surrounded by white space and punctuation, with further splits at certain hyphens and slashes.[1] The tokenisation grammar was applied using recently improved upgrades of the XML tools described in (Thompson et al., 1997) and (Grover et al., 2000).[2] Each

[1] The BioNLP dataset was released pretokenised with splits occurring at whitespace and sentence punctuation. We used this tokenisation as input into our tokeniser which is more fine-grained.

[2] These tools will soon be available under GPL as LT-XML2

Corpus	Total Size	Size of Training Set	Size of Test Set	Entity Markup	Reference
BioNLP	650,720 tokens	540,269 tokens	110,451 tokens	protein (plus DNA, RNA, cell line and cell type)	Collier et al. (2004)
Texas	206,209 tokens	-	-	protein	Bunescu et al. (2005)
PIR	77,528 tokens	-	-	protein	Mani et al. (2005)
Yapex	55,616 tokens	28,201 tokens	27,415 tokens	protein	Franzén et al. (2002)
Bio1	27,476 tokens	-	-	protein (plus DNA, RNA, cell line, cell type, mono-organism, multi-celled organism, virus, sublocation and tissue	Tateisi et al. (2000)

Table 1: Description of different biomedical datasets.

tokenised dataset was subsequently assigned part-of-speech (POS) tags using a Maximum Entropy POS tagger (Curran and Clark, 2003a) trained on the Medpost training data (Smith et al., 2004).

All five corpora which we examined contain documents in the domain of biomedicine, but some of them vary with regard to their type of topic restriction of this broad area. As can be seen in Table 1, the collections also vary in size. The BioNLP corpus is the largest dataset containing 2000 Medline abstracts from the Genia version 3.02 corpus (Ohta et al., 2002). This corpus was constructed by querying Pubmed with the search terms *human*, *transcription factor* and *blood cell*. The BioNLP corpus is marked up for 5 entity classes: protein name, DNA, RNA cell line and cell type. We retained only the protein markup for comparison with other datasets. The BioNLP data is split into a training and a test set. The Texas and the PIR datasets are smaller collections. The former is made up of a total of 750 Medline abstracts containing the word *human*. The latter is a collection of 300 Medline abstracts selected from curated PIR-NREF (Non-Redundant REFerence protein) database entries with no restriction in query. As both the Texas and the PIR datasets contain embedded entities, we trained and tested the tagger on the outermost annotations for such instances. The PIR corpus was annotated following specifically developed guidelines that have also been made publicly available (PIR, 2004). Yapex is one of the smaller corpora with only 200 Medline abstracts manually tagged for protein names. 147 of these abstracts were randomly chosen from a set of Medline abstracts containing the MeSH terms *protein binding*, *interaction* and *molecular*. These make up all of the training and approximately half of the test set. The remaining abstracts in the test set were randomly selected from the Genia corpus (Ohta et al., 2002). Finally, the Bio1 corpus is made up of a random selection of 100 Medline abstracts retrieved from PubMed by querying for *human*, *blood cell* and *transcription factor*. Its content is therefore similar to that of the BioNLP corpus which was collected on the basis of the same search terms. The Bio1 corpus was manually annotated by domain experts for protein names as well as a series of other entities (see Table 1).

3. Protein Tagging Experiments

The experiments we present in this section are designed as follows: we train a Maximum Entropy tagger (Curran and Clark, 2003b) on a portion of a given dataset and then test it on a portion of the same collection (within-dataset) as well as on a different collection (cross-dataset). We use the standard feature set implemented in the tagger, namely prefix and suffix information, morphological and orthographic characteristics, word length, word type, POS tags of current and two previous words as well as context named entity (NE) tags and memory NE tag. The latter is the NE tag that was most recently assigned to the current word. We did not attempt to optimise feature settings for protein name recognition as our aim was not to achieve the best named entity recognition (NER) performance but rather to compare results across different datasets in relation to each other. We show that testing the model on a different corpus yields a severe drop in performance when compared to testing on the same set, and, most importantly, that such a fall is mainly due to discrepancies in the annotation.

3.1. Results

We firstly determine how well the NE tagger is able to recognise protein names in the various within-dataset experiments, i.e. when the training data stems from the same collection as the test data (Table 2). While we use the training and test sets distributed as part of the BioNLP and the Yapex collections for these experiments, we perform 10-fold cross-validation on the Texas, the PIR and the Bio1 corpora. Table 2 shows, for example, that the tagger yields a relatively high precision, recall and F-score when trained and tested on the Bio1 dataset, showing that larger

Within-dataset experiments				
Training Set	Test Set	Precision	Recall	F-Score
BioNLP	BioNLP	64.20%	66.33%	65.25
Texas	Texas	68.84%	46.08%	55.20
PIR	PIR	70.23%	64.37%	67.17
Yapex	Yapex	74.63%	43.69%	55.12
Bio1	Bio1	75.97%	65.74%	70.48

Table 2: Protein name recognition results for within dataset experiments.

and LT-TTT2 at: http://www.ltg.ed.ac.uk

sized training sets do not necessarily guarantee the highest scores. The lower scores of the models trained on the Texas and Yapex datasets also indicate that these collections are likely to contain more internal inconsistencies in their annotation. Interestingly, their low F-scores are due to the low recall scores rather than their precision scores which are comparable with those of the other experiments. The Texas and Yapex models essentially miss many protein names in the test data. We suspect that this could be partly the effect of the inconsistencies in the training data leading the tagger to uncertainty when making a hypothesis. This means that when the tagger is provided with noisy training material, it does not assign the protein tag unless it is fairly confident.

We compare all results from the within-dataset experiments to those of several cross-dataset experiments with the BioNLP data as the test set and the four other datasets as training sets. Each model performs considerably less well on the BioNLP test data than on data from its own collection (see Table 3).

Cross-dataset experiments				
Training Set	Test Set	Precision	Recall	F-Score
Texas	BioNLP	44.05%	24.24%	31.27
PIR	BioNLP	39.81%	42.91%	41.30
Yapex	BioNLP	36.69%	24.29%	29.23
Bio1	BioNLP	46.14%	44.35%	45.22

Table 3: Protein name recognition results for cross-dataset experiments.

However, a number of errors in the cross-dataset experiments are obviously due to unseen tokens in the test data. We therefore ran within- and cross-dataset experiments where all the word features were ignored: the tagger merely relies on word-internal and -external context features but not the identity of the words themselves and is therefore prevented from learning a lexicon of protein names contained in the training data (see Table 4). These are fairer figures for comparing results, as the tagger cannot rely on lexical information from the training set.

Within-dataset experiments (no word-features)				
Training Set	Test Set	Precision	Recall	F-Score
BioNLP	BioNLP	55.27%	30.22%	39.07
Texas	Texas	61.01%	37.18%	46.20
PIR	PIR	63.24%	57.13%	60.03
Yapex	Yapex	64.65%	42.32%	51.15
Bio1	Bio1	69.32%	59.44%	64.00

Table 4: Protein name recognition results for within dataset experiments when ignoring word features.

As Table 5 shows, testing models trained on the Texas, PIR, Yapex and Bio1 data on the BioNLP test data still yields a serious performance fall. Comparing, for example, the results of the Bio1 model in the within- and cross-dataset experiments shows a drop of over 25 points: from 64.00 (training/testing on Bio1) to 38.34 points (testing the Bio1 model on BioNLP). The other models show a similar tendency when tested on the BioNLP data. The average performance drop of all models is 23.7 points.

Cross-dataset experiments (no word-features)				
Training Set	Test Set	Precision	Recall	F-Score
Texas	BioNLP	39.15%	18.59%	25.21
PIR	BioNLP	34.65%	35.98%	35.30
Yapex	BioNLP	31.50%	24.67%	27.67
Bio1	BioNLP	40.69%	36.25%	38.34

Table 5: Protein name recognition results for cross-dataset experiments when ignoring word features.

As could have been anticipated from the previous results, Table 6 shows that combining the different collections for training a model does not lead to better perfomance. When adding the Texas, Yapex and Bio1 datasets to the BioNLP training data, performance actually decreases to 63.58, 64.37 and 64.76 points F-score respectively, compared to when training solely on the BioNLP data, which results in an F-score of 65.25. Adding the PIR data to the BioNLP training data yields a small but negligible improvement. This shows that increasing the amount training data is not useful unless it is consistent.

Combined-dataset experiments				
Training Set	Test Set	Precision	Recall	F-Score
BioNLP+Texas	BioNLP	63.22%	63.94%	63.58
BioNLP+PIR	BioNLP	63.38%	67.34%	65.30
BioNLP+Yapex	BioNLP	65.04%	63.73%	64.37
BioNLP+Bio1	BioNLP	64.91%	64.61%	64.76

Table 6: Protein name recognition results for combined-dataset experiments.

The results of all our experiments illustrate that while a model trained on a specific corpus may perform relatively well on data from the same collection, it may not do so on a new corpus in the same domain. One reason for this dicrepancy in performance are tagging errors caused by annotation inconsistencies across datasets. In the remaining part of the paper, we will present examples of different types of NER errors and examine annotation inconsistencies within and across each of the biomedical datasets.

3.2. Error Analysis

A detailed error analysis of output obtained from cross-dataset experiments illustrates that despite missing protein names as a result of the differences in sub-domain, many tagging errors could have been avoided if the annotation of the different datasets was consistent. For example, we observe many left and right boundary errors, as in the following two examples which occur when testing the Texas and the Bio1 models, respectively, on the BioNLP data:

(1) GOLD: <prot>NF-kappa B complexes</prot>
TAGGED: <prot>NF-kappa B</prot> complexes

(2) GOLD: <prot>human STAT6</prot>
TAGGED: human <prot>STAT6</prot>

There are also errors in the tagging of coordinated NPs as well as abbreviations. The following two extracts are examples of such errors that occurred in the output of the Yapex and the PIR model:

(3) GOLD: <prot>HSV-1 ICP0 and CMV IE1 proteins</prot>
TAGGED: HSV-1 <prot>ICP0</prot> and <prot>CMV IE1</prot> proteins

(4) GOLD: <prot>interleukin 2 (IL 2)</prot>
TAGGED: <prot>interleukin 2</prot> (<prot>IL 2</prot>)

Moreover, there are more false negatives in the case of general protein names (annotated in the BioNLP gold standard) if they were not annotated in the training dataset, as for example in the Texas data:

(5) GOLD: predominant <prot>cellular proteins</prot>
TAGGED: predominant cellular proteins

These are a selection of error types caused by annotation inconsistencies across datasets which are investigated in the following section.

4. Annotation Inconsistencies

Error analysis showed that inconsistencies in the annotation of protein names across datasets are a main cause of NER errors in the cross-dataset experiments. In this section, we present different types of inconsistencies which exist not only across but also within datasets. We mainly focus on protein name boundaries, coordination, abbreviations as well as general protein names and provide relevant examples.[3]

4.1. Protein Name Boundaries

Boundary inconsistencies of entities are one major difference in the annotation across datasets. For example, the word *protein* following a protein name (i.e. "X protein") is always annotated as part of the entity in the PIR and Bio1 datasets.[4] Similarly, it tends to be included, although not consistently, in the annotation of the BioNLP dataset (in 90.5% of cases). Both the Texas and Yapex datasets in turn contain some annotations where the word *protein* is included (47.4% and 62.7%, respectively) and some where it is not (52.6% and 37.3%, respectively) and therefore present very contradictory information to the learner. The Texas and Yapex datasets show a more consistent treatment of the word *complex*, which is never included in the annotation at the end of a protein name. However, the word *complex* in that same position is always annotated in the PIR data and in 96.4% of cases in the BioNLP data. In the BioNLP data, for example, the string *AP-1 complex* is in all but one cases annotated as a full protein name. The Bio1 dataset only contains very few protein names followed by the word *complex*, which is either included (72%) or not (28%). One type of entity where the annotations are consistent across datasets at least in the majority of cases is when a protein name ends in the word *receptor* which is generally considered part of the named entity. However, only the PIR and Bio1 datasets are completely consistent in their markup. All other datasets also contain annotations where the entity does not span the word *receptor* at the end. Their number is smallest for the BioNLP data at 3% but larger for the Texas and Yapex datasets at 17.3% and 17.6%, respectively.

All previous examples examined right boundaries of protein names. Left boundaries of protein names vary equally across datasets particularly with regard to the decision of whether to consider an organism preceding a protein name as part of the entity or not. Organisms are mostly included in the protein name in the annotation of the BioNLP and the PIR data as in the following example:

(6) . . . <prot>human CD14</prot> . . .

Nevertheless, we found a small number of inconsistencies in both datasets. The Texas and the Yapex annotations do not in the majority of cases include the organism in the protein entity even if it is specified as part of the protein name indicated by an abbreviation or acronym, e.g.:

(7) . . . human <prot>leukocyte antigen</prot> (<prot>HLA</prot>) . . .

Similarly, organisms are also not annotated as part of the protein name in the Bio1 data but as a "multi-cell source" entity preceding a protein entity.

4.2. Coordinated Protein Named Entities

Regarding coordination of type "X and/or Y proteins", the actual protein names *X* and *Y* are annotated as two separate entities in the Texas and the Yapex datasets, e.g. see Example (8) taken from the Texas data:

(8) . . . <prot>Rad51</prot> or <prot>Rad52</prot> proteins . . .

The annotation of this type of coordinated NP differs in all other collections. There the word *proteins* which modifies all coordinated members of the NP is annotated as part of the last protein name, e.g. see Example (9) taken from the PIR data:

(9) . . . <prot>hRad54</prot> and <prot>hRad54B proteins</prot> . . .

In the BioNLP test data we also detected a number of cases where the entire coordinated NP with all protein members was annotated as one protein name entity.

4.3. Abbreviations and Acronyms

In the PIR data, a protein name abbreviation is only annotated as a separate entity in case it precedes additional information in the brackets. Generally, full protein names and their abbreviations or acronyms which co-occur in the

[3] The examples given in this section contain a common XML markup and not the original markup convention of each dataset.

[4] We do not distinguish between protein names ending in the word protein and those that are modified by the word protein.

text in either order are marked up as one entity in this collection, e.g.:

(10) ...<prot>ATR (anthrax toxin receptor)</prot> ...

(11) ...<prot>transforming growth factor alpha (TGF-alpha)</prot> ...

In all the other datasets, full names as well as abbreviations and acronyms are mostly considered as separate protein names unless they are part of a larger protein name. So for example, the string *interleukin 2 (IL-2)* contains the protein names *interleukin 2* and *IL-2* as shown in Example (12).

(12) ...<prot>interleukin 2</prot> (<prot>IL-2 </prot>) ...

We have found some inconsistencies within several of the data sets. For example, the string *interleukin 2 (IL-2)* in Example (12) taken from the Bio1 corpus also occurs in the same collection annotated as one protein name in a similar fashion to in the PIR data.

4.4. General Protein Names

Proteins are also frequently described in more general terms, such as *cellular protein*, *regulatory protein* or *integral membrane protein*. These general expressions are deliberately annotated as entities in the BioNLP and the PIR datasets. Conversely, they are not considered as protein entities in the other datasets. For example the string *integral membrane protein* is marked up completely differently in the BioNLP and the Yapex data (see Examples (13) and (14), respectively).

(13) ...<prot>integral membrane protein</prot> ...

(14) ...integral membrane protein ...

In the BioNLP and the PIR data, such general protein descriptions are even annotated if they refer to more than one specific protein, e.g. *regulatory proteins*. It becomes clear that the annotators working on those two corpora had or were provided with a very different concept of what constitutes a named entity compared to those employed to mark up the other data sets.

5. Discussion and Conclusion

We discussed a series of inconsistencies that exist across and even within different datasets which cause boundary errors and increase the number of false positives. This means that increasing the size of the training data by adding new annotated data is never beneficial unless the annotations of individual collections happen to be similar enough. Even then, inconsistencies within a dataset are detrimental to the tagger's performance. Considering the increasing amount of work dedicated to information extraction from biomedical publications, we believe that the creation of standard annotation guidelines for various entities of interest will not only improve recognition performance but also encourage the shareability of resources (see also Cohen et al. (2005a) on features that make a corpus useful). The experiments we have presented have singled out specific annotation issues that we believe should receive priority attention in any new annotation project. When comparing the various annotation schemes of all the biomedical corpora we have also observed slight differences in the definition of specific entity types. We believe that future efforts in biomedical corpus annotation will become more attractive if they follow the standardised definitions and guidelines.

6. Acknowledgements

We would like to thank Barry Haddow for his comments and assistance. This work was performed as part of the SEER project which is supported by a Scottish Enterprise Edinburgh-Stanford Link Grant (R36759).

7. References

Ann Bies, Seth Kulick, and Mark Mandel. 2005. Parallel entity and treebank annotation. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 21–28, Ann Arbor, Michigan, USA. Association for Computational Linguistics.

Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun Kumar Ramani, and Yuk Wah Wong. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2):139–155.

Kevin Bretonnel Cohen, Lynne Fox, Philip V. Philip Ogren, and Lawrence Hunter. 2005a. Corpus design for biomedical natural language processing. In *Proceedings of the ACL-ISMB workshop on linking biological literature, ontologies and databases*, pages 38–45, Detroit, Michigan, USA. Association for Computational Linguistics.

Kevin Bretonnel Cohen, Lynne Fox, Philip V. Philip Ogren, and Lawrence Hunter. 2005b. Empirical data on corpus design and usage in biomedical natural language processing. In *Proceedings of AMIA 2005*, pages 156–160.

Nigel Collier, Patrick Ruch, and Adeline Nazarenko, editors. 2004. *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)*, Geneva, Switzerland.

James R. Curran and Stephen Clark. 2003a. Investigating GIS and smoothing for maximum entropy taggers. In *Proceedings of the 11th Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL'03)*, pages 91–98, Budapest, Hungary.

James R. Curran and Steven Clark. 2003b. Language independent NER using a maximum entropy tagger. In *Proceedings of the 7th Conference on Natural Language Learning*, Edmonton, Canada.

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(1-3):49–61.

Claire Grover, Colin Matheson, Andrei Mikheev, and Marc Moens. 2000. LT TTT - a flexible tokenisation tool. In

Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000), Athens, Greece.

Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Antonio Valencia. 2004. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bionformatics*, 6(Suppl 1:S1).

Inderjeet Mani, Zhangzhi Hu, Seok Bae Jang, Ken Samuel, Matthew Krause, Jon Phillips, and Cathy H. Wu. 2005. Protein name tagging guidelines: lessons learned. *Comparative and Functional Genomics*, 6(1-2):72–76.

Tomoko Ohta, Yuka Tateisi, and Jin-Dong Kim. 2002. The GENIA corpus: An annotated research abstract corpus in molecular biology domain. In *Proceedings of the Human Language Technology Conference*, pages 73–77, San Diego, California, USA.

PIR, 2004. *Guidelines for Protein Name Tagging Version 2.0*. Protein Information Resource, Georgetown University Medical Center and Department of Linguistics Georgetown University, April.

Lawrence H. Smith, Tom C. Rindflesch, and W. John Wilbur. 2004. Medpost: a part-of-speech tagger for biomedical text. *Bioinformatics*, 20(14):2320–2321.

Yuka Tateisi, Tomoko Ohta, Nigel Collier, Chikashi Nobata, and Jun-ichi Tsujii. 2000. Building an annotated corpus in the molecular-biology domain. In *Proceedings of COLING'2000 Workshop on Semantic Annotation and Intelligent Content*, Luxemburg.

Henry S. Thompson, Richard Tobin, and David McKelvie, 1997. *LT XML. Software API and toolkit for XML processing*. URL: `http://www.ltg.ed.ac.uk/software/`.