Empirical data on corpus design and usage in biomedical natural language processing

K. Bretonnel Cohen
Center for Computational Pharmacology
U. of Colorado School of Medicine
kevin.cohen@gmail.com

Lynne Fox
Denison Library
U. of Colorado Health Sciences Center
lynne.fox@uchsc.edu

Philip V. Ogren
Center for Computational Pharmacology
U. of Colorado School of Medicine
philip.ogren@uchsc.edu

Lawrence Hunter
Center for Computational Pharmacology
U. of Colorado School of Medicine
larry.hunter@uchsc.edu

Abstract

This paper describes the designs of six publicly available biomedical corpora. We then present usage data for the six corpora. We show that corpora that are carefully annotated with respect to structural and linguistic characteristics and that are distributed in standard formats are more widely used than corpora that are not. These findings have implications for the design of the next generation of biomedical corpora.

1 Introduction

A small number of data sets for evaluating the performance of biomedical language processing systems on a small number of task types have been made publicly available by their creators[1]. From a biological perspective, a number of these corpora (PDG, GENIA, Medstract, Yapex) are exceptionally well curated. From the perspective of system evaluation, a number of these corpora (Wisconsin, GENETAG) are very well designed, with large numbers of both positive and negative examples for system training and testing. Despite the positive attributes of all of these corpora, they vary widely in their external usage rates: some of them have been found very useful in the natural language processing community outside of the labs that created them, as evinced by their high rates of usage in system construction and evaluation in the years since they have been released. In contrast, others have seen little or no use in the community at large. These data sets provide us with an opportunity to evaluate the consequences of a variety of approaches to biomedical corpus construction. We examine these corpora with respect to a number of design features and other characteristics, and look for features that characterize widely used—and infrequently used—corpora.

1.1 Overview of corpus designs

The unusual formats of two of the corpora probably contribute to their low external usage rates, and therefore they merit special discussion. The PDG corpus was built at the very beginning of the involvement of the computational biology community in text data mining efforts. Its eventual public distribution was not anticipated at the time of its construction, and it is the least annotated of the six corpora. The data is made available as an HTML file, which necessitates stripping formatting tags before use. The data is in two sections within the single document. The two sections represent two Drosophila signalling systems.

The example below shows a representative datum. Proteins that are involved in the relation are indicated on the *Proteins* line. Proteins that are not (e.g. *cdk4* in the example) are not annotated. All text is normalized to lower-case. Unless an entry contains more than one sentence, the sentence-final punctu-

[1] These are the corpora described in Blaschke et al. 1999, which we refer to as the Protein Design Group (PDG) corpus; Craven and Kumlein 1999, which we refer to as the University of Wisconsin corpus; Collier et al. 1999, the GENIA corpus; Pustejovsky et al. 2002, the Medstract corpus; Franzén et al. 2002, the Yapex corpus; and Tanabe et al. 2005, originally the BioCreative Task 1A data set, now known as the GENETAG corpus.

ation is normalized away. Protein names are also normalized to some extent; for example, *cyclin d2* appears in the annotation as *cyclin D*. Linguistically, the data is unannotated. However, it should be noted that from a biological perspective, the data is highly relevant and of exceptionally high quality.

```
MED 97322239:
actions:  activates;
Proteins: cdk2; cyclin D;
cyclin d2 activates cdk2 in
preference to cdk4 in human breast
epithelial cells
```

Although the bulk of the data in the file is in this format, another format is used elsewhere in the file, requiring considerable data manipulation.

The Wisconsin data is especially notable for its large size—at over a million and a half words, it is the largest of the corpora. The data was assembled from literature references in publicly available databases. The explicit philosophy of the construction process is to make use of "lightly annotated" freely available data. The semantic annotation is automatic and is based on metadata from the original data sources. The linguistic annotation, which consists of tokenization, part-of-speech tagging, and shallow parsing, is automatic and is not manually curated[2].

The following example shows a representative positive datum[3]. The first line (*"PIGA(-) cells... hemoglobinuria.. ")*) contains the text. The second line contains the entities involved in the relation (in this example, *[PIGA,Paroxysmal nocturnal hemoglobinuria]* (a gene/disease association from OMIM)). Note that their forms are normalized, making it awkward to map from this annotation to the raw text. The next line (*[0,12]*)contains the indices of the base phrases that contain those entities per the shallow parser output that follows it. Note that punctuation has been normalized away completely from this representation, including the crucial *(-)* which indicates that this is a knocked-out gene.

```
"PIGA(-) cells had no growth
advantage, suggesting that other
factors are needed for their
clonal dominance in patients with
paroxysmal nocturnal
hemoglobinuria.. "
[PIGA,Paroxysmal nocturnal hemoglobinuria]
[0,12]

0 NP_SEGMENT:GENE piga{UNK:GENE} cells{N}
1 VP_SEGMENT had{V}
2 NP_SEGMENT no{ADJ} growth{UNK} advantage{N}
```

In addition to removing punctuation, the tokenization process also joins together the elements of multi-word terms, e.g. *amino_acid, nuclear_membrane,* and *because_of*. This strategy addresses some problems, but also makes it difficult to relate annotation to the raw text. Note that from a system evaluation perspective, this corpus is very well designed.

The formats of the other corpora are relatively standard and require little discussion. Three of them (GENIA, Medstract, and Yapex) are in XML, and one (GENETAG) is in the familiar Brill tagger format.

Table 1: Name, date, genre, and size for the six corpora. Size is in words.

Name	date	genre	size
PDG	1999	Sentences	10,291
Wisconsin	1999	Sentences	1,529,731
GENIA	1999	Abstracts	432,560
Medstract	2001	Abstracts	49,138
Yapex	2002	Abstracts	45,143
GENETAG	2004	Sentences	342,574

2 Materials and methods

Table 1 lists the biomedical corpora available as of Spring 2005[4].

For each one, it gives its release date (or the year of the corresponding publication), the genre of the contents of the corpus, and the size of the corpus[5].

The left-hand side of Table 2 lists the data sets and, for each one, indicates the lower-level language processing problems that it could be applied to, either as a source of training data or for evaluating sys-

[2]It is due to this lack of curation that we do not indicate this data as being applicable to the sentence segmentation, tokenization, or POS-tagging tasks in Table 2.

[3]We truncated the shallow parser output due to space considerations.

[4]We omit text collections from our discussion. By *text collection* we mean textual data sets that may include metadata about documents, but do not contain mark-up of the document contents.

[5]Published descriptions of the corpora don't generally give size in words, so this data is based on our own counts.

Table 2: Low- and high-level tasks to which the six corpora are applicable. SS is sentence segmentation, T is tokenization, and POS is part-of-speech tagging. EI is entity identification, IE is information extraction, and C is coreference resolution.

Name	SS	T	POS	EI	IE	C
PDG				•	•	
Wisconsin				•	•	
GENIA	•	•	•	•		
Medstract				•		•
Yapex				•		
GENETAG				•		

tems that perform these tasks. We considered here sentence segmentation, word tokenization, and part-of-speech (POS) tagging.

The right-hand side of Table 2 shows the higher-level tasks to which the various corpora can be applied. We considered here entity identification, information extraction, and coreference resolution. These tasks are directly related to the types of semantic annotation present in each corpus. The three EI-only corpora (GENIA, Yapex, GENETAG) are annotated with semantic classes of relevance to the molecular biology domain. In the case of the Yapex and GENETAG corpora, this annotation uses a single semantic class, roughly equivalent to the gene or gene product. In the case of the GENIA corpus, the annotation reflects a more sophisticated ontology. The Medstract corpus uses multiple semantic classes, including *gene*, *protein*, *cell type*, and *molecular process*. In all three cases, the semantic annotation was carefully curated, and in one (GENETAG) it includes alternative analyses.

Two of the corpora (PDG, Wisconsin) are indicated in Table 2 as being applicable both to EI and to IE tasks. From a biological perspective, the PDG corpus has exceptionally well-curated positive examples. From a language processing perspective, it is unannotated. For each sentence, the entities are listed, but their locations in the text are not indicated, making them applicable to some definitions of the entity identification task but not others. The Wisconsin corpus contains both positive and negative examples. For each example, entities are listed in a normalized form, but without clear pointers to their locations in the text, making this corpus similarly difficult to apply to many definitions of the entity identification task.

The Medstract corpus is unique at this time in being annotated with coreferential equivalence sets.

All six corpora draw on the same subject matter domain—biomedicine—but they vary widely with respect to their level of semantic restriction within that relatively broad category. One (GENIA) is restricted to the subdomain of human blood cell transcription factors. Another (Yapex) combines data from this domain with abstracts on protein binding in humans. The GENETAG corpus is considerably broader in topic, with all of PubMed/MEDLINE serving as a potential data source. The Medstract corpus contains biomedical material not apparently related to molecular biology. The PDG corpus is drawn from a very narrow subdomain on protein-protein interactions. The Wisconsin corpus is composed of data from three separate narrow subdomains: protein-protein interactions, subcellular localization of proteins, and gene/disease associations.

Table 3 shows the number of systems *built outside of the lab that created the corpus* that used each of the data sets described in Tables 1 and 2. The counts in this table reflect work that actually used the datasets, versus work that cites the publication that describes the corpus but doesn't actually use it. We assembled the data for these counts by consulting with the creators of the data sets and by doing literature searches. If a system is described in multiple publications, we count it only once, so the number of systems is slightly smaller than the number of publications.

3 Results

Even without examining the external usage data, two points are immediately evident from Tables 1 and 2:

- Only one of the currently publicly available corpora is suitable for evaluating performance on basic preprocessing tasks.
- The currently publicly available corpora include a very limited range of genres: only abstracts and roughly sentence-sized inputs are represented.

Table 3: External usage rates. The *systems* column gives the count of the number of systems that actually used the dataset, as opposed to publications that cited the paper but did not use the data itself. *Age* is in years as of 2005.

Name	age	systems
GENIA	6	21
GENETAG	1	8
Yapex	3	6
Medstract	4	3
Wisconsin	6	1
PDG	6	0

Examination of Table 3 makes another point immediately clear. Some corpora see considerable external use, and others do not. We now consider a number of design features and other characteristics of these corpora that might explain these groupings.

3.1 Effect of age

We considered the possibility that the length of time that a corpus has been available determines the number of external uses. Table 3 shows clearly that this is not the case. The age of the PDG, Wisconsin, and GENIA data is the same, but the usage rates are considerably different—the GENIA corpus has been much more widely used. The GENETAG corpus is the newest, but has a relatively high usage rate. Usage of a corpus is determined by factors other than the length of time that it has been available.

3.2 Effect of size

We considered the possibility that size might be the determinant of the amount of external use—perhaps smaller corpora simply do not provide enough data to be helpful in the development and validation of learning-based systems. We found that size does not determine use. The Yapex corpus is one of the smallest, but has achieved fairly wide usage. The Wisconsin corpus is the largest, but has a very low usage rate.

3.3 Effect of structural and linguistic annotation

We expected that the corpus with the most structural and linguistic annotation would have the highest usage rate[6]. The extent to which this is true is not clear. The GENIA corpus is the only one with curated structural and POS annotation, and it has the highest usage rate. This is consistent with our a priori expectation. On the other hand, the Wisconsin corpus could be considered the most "deeply" linguistically annotated, since it has both POS annotation and—unique among the various corpora—shallow parsing. It nevertheless has a very low usage rate.

However, the comparison is not clearcut, since both the POS tagging and the shallow parsing in the Wisconsin corpus are fully automatic and not manually corrected. (Additionally, the shallow parsing and the tokenization on which it is based are somewhat idiosyncratic.) It *is* clear that the Yapex corpus has relatively high usage despite the fact that it is, from a structural and linguistic perspective, unannotated (it is marked up for entities only, and nothing else.) To our surprise, structural and linguistic annotation do not appear to uniquely determine usage rate.

3.4 Effect of format

Annotation format has a large effect on usage. It bears repeating that these six different corpora are distributed in six different formats—even the presumably simple task of populating the *Size* column in Table 1 required writing six separate scripts to parse the various data files. The two low-usage corpora are annotated in remarkably unique formats. In contrast, the three more widely used corpora are distributed in relatively more common formats. Three of them (GENIA, Medstract, and Yapex) are distributed in XML, and one of them (GENIA) offers a choice for POS tagging information between two well-known formats. The fourth (GENETAG) is distributed in the widely used Brill tagger format.

3.5 Effect of semantic annotation

The data in Tables 2 and 3 are consistent with the hypothesis that semantic annotation predicts usage. The claim would be that corpora that are built specifically for entity identification purposes are more widely used than corpora of other types, presumably

[6]By *structural annotation* we mean tokenization and sentence segmentation, and by *linguistic annotation* we mean POS tagging and shallow parsing.

due to a combination of the importance of the EI task as a prerequisite to a number of other important applications and the fact that EI is still an unsolved problem in the biomedical domain. However, there are large differences in the usage rates of the three EI corpora, suggesting that semantic annotation is not the only relevant design feature. *If* semantic annotation determines usage, then one would predict a reduction in the use of all three of the EI-only corpora once the EI problem is solved, unless their semantic annotations are extended in new directions.

3.6 Effect of semantic domain

We considered that the extent of restriction of the semantic domain might determine usage. It does not: both the low-use and high-use groups of corpora contain at least one highly restricted domain (GENIA in the high-use group, and PDG in the low-use group) and one broader domain (GENETAG in the high-use group, and Wisconsin in the lower-use group).

4 Discussion

Corpus construction efforts can consume large amounts of time and resources. Corpora that are not widely used represent considerable losses of intellectual, as well as literal, capital. Our data suggests that future corpora can help ensure their usability and usefulness by choosing standard formats for annotation and distribution, and by including high-quality annotation of structural and linguistic characteristics of their contents. Furthermore, the investment in the extant low-usage corpora can be recovered by curating their current annotations, adding curated structural and linguistic annotation where they are absent, and standardizing their annotation formats. Three corpora that follow many of the maxims of corpus construction with respect to format and annotation—the PennBioIE (Kulick et al. 2004), the MedTag (Smith et al. 2005), and SICS's FetchProt corpora—have recently or will soon become publicly available, and we predict high usage for them.

5 Acknowledgments

We gratefully acknowledge help from C. Blaschke, M. Craven, K. Franzén, T. Gibson, L. Hirschman, A. Morgan, S. Leach, T. Ohta, L. Tanabe, and Y. Tateisi.

References

Blaschke, Christian; Miguel A. Andrade; Christos Ouzounis; and Alfonso Valencia. 1999. Automatic extraction of biological information from scientific text: protein-protein interactions. *Proceedings of ISMB-99*, pp. 60-67. AAAI Press.

Collier, Nigel, Hyun Seok Park, Norihiro Ogata, Yuka Tateisi, Chikashi Nobata, Takeshi Sekimizu, Hisao Imai and Jun'ichi Tsujii. 1999. The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers. *Proceedings of the European Association for Computational Linguistics (EACL 1999)*.

Craven, Mark; and Johan Kumlein. 1999. Constructing biological knowledge bases by extracting information from text sources. *Proceedings of ISMB-99*, pp. 77-86. AAAI Press.

Franzén, Kristofer; Gunnar Eriksson; Fredrik Olsson; Lars Asker Per Lidin; and Joakim Cöster. 2002. Protein names and how to find them. *International Journal of Medical Informatics*, 67(1-3), pp. 49-61.

Kulick, Seth; Ann Bies; Mark Liberman; Mark Mandel; Ryan McDonald; Martha Palmer; Andrew Schein; and Lyle Ungar. 2004. Integrated annotation for biomedical information extraction. *Proceedings of the HLT/NAACL workshop BioLink 2004: Linking biological literature, ontologies and databases*, pp. 61-68.

Pustejovsky, J.; J. Castanño; R. Saurı'; A. Rumshisky; J. Zhang; and W. Luo. 2002. Medstract: creating large-scale information servers for biomedical libraries. *Proceedings of the workshop on natural language processing in the biomedical domain*, pp. 85-92. Association for Computational Linguistics.

Smith, Lawrence H.; Lorraine Tanabe; Thomas Rindflesch; and W. John Wilbur. 2005. MedTag: a collection of biomedical annotations. *Proceedings of the ACL-ISMB workshop on linking biological literature, ontologies and databases: mining biological semantics*, pp. 32-37.

Tanabe, Lorraine; 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.