Refactoring Corpora

Helen L. Johnson
Center for Computational Pharmacology
U. of Colorado School of Medicine
helen.johnson@uchsc.edu

William A. Baumgartner, Jr.
Center for Computational Pharmacology
U. of Colorado School of Medicine
william.baumgartner@uchsc.edu

Martin Krallinger
Protein Design Group
Universidad Autónoma de Madrid
martink@cnb.uam.es

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

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

Abstract

We describe a pilot project in semi-automatically refactoring a biomedical corpus. The total time expended was just over three person-weeks, suggesting that this is a cost-efficient process. The refactored corpus is available for download at http://bionlp.sourceforge.net.

1 Introduction

Cohen et al. (2005) surveyed the usage rates of a number of biomedical corpora, and found that most biomedical corpora have not been used outside of the lab that created them. Empirical data on corpus design and usage suggests that one major factor affecting usage is the format in which it is distributed.

These findings suggest that there would be a large benefit to the community in refactoring these corpora. *Refactoring* is defined in the software engineering community as altering the internal structure of code without altering its external behavior (Fowler et al., 1999). We suggest that in the context of corpus linguistics, refactoring means changing the *format* of a corpus without altering its *contents*, i.e. its annotations and the text that they describe. The significance of being able to refactor a large number of corpora should be self-evident: a likely increase in the use of the already extant publicly available data for evaluating biomedical language processing systems, without the attendant cost of repeating their annotation.

We examined the question of whether corpus refactoring is practical by attempting a proof-of-concept application: modifying the format of the Protein Design Group (PDG) corpus described in Blaschke et al. (1999) from its current idiosyncratic format to a stand-off annotation format (WordFreak[1]) and a GPML-like (Kim et al., 2001) embedded XML format.

2 Methods

The target WordFreak and XML-embedded formats were chosen for two reasons. First, there is some evidence suggesting that standoff annotation and embedded XML are the two most highly preferred corpus annotation formats, and second, these formats are employed by the two largest extant curated biomedical corpora, GENIA (Kim et al., 2001) and BioIE (Kulick et al., 2004).

The PDG corpus we refactored was originally constructed by automatically detecting protein-protein interactions using the system described in Blaschke et al. (1999), and then manually reviewing the output. We selected it for our pilot project because it was the smallest publicly available corpus of which we were aware. Each block of text has a deprecated MEDLINE ID, a list of actions, a list of proteins and a string of text in which the actions and proteins are mentioned. The structure and contents of the original corpus dictate the logical steps of the refactoring process:

1. Determine the current PubMed identifier, given the deprecated MEDLINE ID. Use the PubMed identifier to retrieve the original abstract.

[1] http://venom.ldc.upenn.edu/_resources/info/wordfreak_ann.html

Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 06, pages 116–117,
New York City, June 2006.

2. Locate the original source sentence in the title or abstract.
3. Locate the "action" keywords and the entities (i.e., proteins) in the text.
4. Produce output in the new formats.

Between each file creation step above, human curators verify the data. The creation and curation process is structured this way so that from one step to the next we are assured that all data is valid, thereby giving the automation the best chance of performing well on the subsequent step.

3 Results

The refactored PDG corpus is publicly available at http://bionlp.sourceforge.net. Total time expended to refactor the PDG corpus was 122 hours and 25 minutes, or approximately three person-weeks. Just over 80% of the time was spent on the programming portion. Much of that programming can be directly applied to the next refactoring project. The remaining 20% of the time was spent curating the programmatic outputs.

Mapping IDs and obtaining the correct abstract returned near-perfect results and required very little curation. For the sentence extraction step, 33% of the corpus blocks needed manual correction, which required 4 hours of curation. (Here and below, "curation" time includes both visual inspection of outputs, and correction of any errors detected.) The source of error was largely due to the fact that the sentence extractor returned the best sentence from the abstract, but the original corpus text was sometimes more or less than one sentence.

For the protein and action mapping step, about 40% of the corpus segments required manual correction. In total, this required about 16 hours of curation time. Distinct sources of error included partial entity extraction, incorrect entity extraction, and incorrect entity annotation in the original corpus material. Each of these types of errors were corrected.

4 Conclusion

The underlying motivation for this paper is the hypothesis that corpus refactoring is practical, economical, and useful. Erjavec (2003) converted the GENIA corpus from its native format to a TEI P4 format. They noted that the translation process brought to light some previously covert problems with the GENIA format. Similarly, in the process of the refactoring we discovered and repaired a number of erroneous entity boundaries and spurious entities.

A number of enhancements to the corpus are now possible that in its previous form would have been difficult at best. These include but are not limited to performing syntactic and semantic annotation and adding negative examples, which would expand the usefulness of the corpus. Using revisioning software, the distribution of iterative feature additions becomes simple.

We found that this corpus could be refactored with about 3 person-weeks' worth of time. Users can take advantage of the corrections that we made to the entity component of the data to evaluate novel named entity recognition techniques or information extraction approaches.

5 Acknowledgments

The authors thank the Protein Design Group at the Universidad Autónoma de Madrid for providing the original PDG protein-protein interaction corpus, Christian Blaschke and Alfonso Valencia for assistance and support, and Andrew Roberts for modifying his jTokeniser package for us.

References

Christian Blaschke, Miguel A. Andrade, and Christos Ouzounis. 1999. Automatic extraction of biological information from scientifi c text: Protein-protein interactions.

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

Tomaz Erjavec, Yuka Tateisi, Jin-Dong Kim, and Tomoko Ohta. 2003. Encoding biomedical resources in TEI: the case of the GENIA corpus.

Martin Fowler, Kent Beck, John Brant, William Opdyke, and Don Roberts. 1999. *Refactoring: improving the design of existing code*. Addison-Wesley.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, Hideki Mima, and Jun'ichi Tsujii. 2001. Xml-based linguistic annotation of corpus. In *Proceedings of The First NLP and XML Workshop*, pages 47–53.

S. Kulick, A. Bies, M. Liberman, M. Mandel, R. McDonald, M. Palmer, A. Schein, and L. Ungar. 2004. Integrated annotation for biomedical information extraction. *Proceedings of the HLT/NAACL*.