A Tool for the Automatic and Manual Annotation of Biomedical Documents

Anália Lourenço[1], Sónia Carneiro[1], Rafael Carreira[2], Miguel Rocha[2], Isabel Rocha[1], Eugénio Ferreira[1]
[1] IBB - Institute for Biotechnology and Bioengineering, Center of Biological Engineering
[2] Department of Informatics / CCTC
University of Minho
Campus de Gualtar, 4710-057 Braga – PORTUGAL
{analia,soniacarneiro,
ecferreira,irocha}@deb.uminho.pt,
{rafaelcc,mrocha}@deb.uminho.pt

Abstract

The techniques developed within the field of Biomedical Text Mining (BioTM) have been mainly tested and evaluated over a set of known corpora built by a few researchers with a specific goal or to support scientific competitions. The generalized use of BioTM software therefore requires that an enlarged set of corpora is made available covering a wider range of biomedical research topics. This work proposes a software tool that facilitates the task of building a BioTM corpus by providing a user-friendly and interoperable tool that allows both automatic and manual annotation of biomedical documents (supporting both abstracts and full text). This tool is also integrated in a more comprehensive BioTM framework.

1 Introduction

Semantic annotation, sometimes called concept matching in the biomedical literature, is the process of mapping phrases within a source text to distinct concepts defined by domain experts.

Traditionally, such annotation was exclusively manual. However, the growing scientific publication rate, the continuous evolving of biological terminology and the more complex analysis requirements brought by systems-level approaches urge for automated curation processes (Ananiadou et al., 2006; Natarajan et al., 2005; Erhardt et al., 2006).

The research field of BioTM emerged from this need and has been providing for helpful computerised approaches. In particular, Biomedical Named Entity Recognition (BioNER), the field that deals with the unambiguous identification of named entities (such as names of genes, proteins, gene products, organisms, drugs, chemical compounds, etc.), is the key step for accessing and integrating the information stored in the literature (Zweigenbaum et al., 2007; Jensen et al., 2006; Natarajan et al., 2005).

Techniques for term identification are becoming widely used in biomedical research. Lexical resources (Fundel and Zimmer, 2006; Mukherjea et al., 2004; Kou et al., 2005; Muller et al., 2004) and rule-based systems (Hu et al., 2005; Hanisch et al., 2005) deliver some degree of automation. On the other hand, Machine Learning contributions (Okazaki and Ananiadou, 2006; Kou et al., 2005; Shi and Campagne, 2005; Yeganova et al., 2004; Sun et al., 2006) address issues like term novelty, synonymy (including term variants and abbreviations) and homonymy.

Despite current achievements, technique development and usage are constrained by the limited availability of high-quality training corpora. In fact, at this point, biomedical annotated corpora represent a bottleneck in the development of BioTM software. Existing approaches cannot be extended without the production of corpora, conveniently validated by domain experts.

In this work, a contribution to tackle this matter is provided, with the development of a novel interoperable and user-friendly software application that supports manual curation of biomedical documents. The proposed software implements a workflow where a biomedical corpus is automatically annotated based on a specialised dictionary. The discovered biomedical concept output is then directed into a manual curation stage, and finally a high-quality biomedical annotated corpus is made available.

Both the automatic and manual annotation tasks are envisioned to be flexible, allowing the tagging of many biological entity classes and the creation and use of different dictionaries, extracted from major biomedical databases. Although we have our own annotation schema, the software is expected to be useful within other domains which have domain-specific resources available. In other words, if a new annotation schema is defined and the dictionary builders cope with it, both automatic and manual annotation are granted.

The remainder of this paper starts by placing annotation tools within BioTM scenario, establishing basic requirements and identifying related work. The enumeration of the software development aims follows. Next, the main features of the proposed software application are discussed, namely the creation of particular dictionaries, the default annotation schema, the automatic annotation module and user-friendly manual annotation environment. Final remarks provide an overall perspective of the work and identify new features.

2 The Role of Annotation Tools in BioTM

Emerging efforts in BioTM agree on considering manually annotated biomedical corpora as priceless resources (Kim et al., 2008; Kim et al., 2003). Many researchers openly contribute and disseminate annotated corpora such as GENIA (Kim et al., 2003), PennBioIE (Kulick S et al., 2004) or GENETAG (Tanabe et al., 2005). Also, there are datasets coming from knowledgeable challenges such as BioCreAtive[1]. Yet, adaptation of available resources to new problems (real-world scenarios) usually requires substantial efforts, since they have been designed to meet a particular aim and tend not to comply with any common data format.

The construction of a new corpus implies the laborious and time-consuming manual collection and annotation of a significant number (typically hundreds) of documents. It is not straightforward to gather, organise and annotate a valuable set of documents. On the one hand, the set of documents has to be representative of the domain it is supposed to describe, i.e., it has to embrace the terminological trends that characterise the domain, while establishing a contrast towards other domains. On the other hand, annotation has to be as comprehensible and consensual as possible. According to a given annotation schema, different annotators should be able to agree, producing similar outputs. Otherwise, either the annotation schema is not able to reflect the domain conveniently, or the domain requires further annotation rules that prevent contradicting or misleading outputs.

It is not reasonable to acknowledge the need for corpora without devising computational annotation tools. There exist several manual text annotation tools for creating annotated corpora. General-purpose annotation tools such as Callisto[2], WordFreak[3](Morton and LaCivita, 2003), the General Architecture for Text Engineering (GATE[4]) (Cunningham et al., 2002) and MMAX2[5] are references in the area. However, these tools present limited flexibility and its 'out of the box' usage often demands expert programming skills.

Although offering customisable tasks (for example, a simple annotation schema can be defined with an XML DTD), these tools do not offer any support for biology-related natural language processing. Dedicated tools such as POS taggers, parsers and named entity recognisers are becoming widely available and it would be desirable to include them into annotation tools.

Tools should support semantic annotation by hand and some form of automatic annotation (using available resources such as dictionaries, ontologies, templates or user-specified rules). Moreover, by supporting both syntactic and semantic annotation, a wide variety of annotation schemas can be defined and used. New annotation tasks can be built without writing new software or creating specialised configuration files.

3 Development Aims

[1] http://biocreative.sourceforge.net/

[2] http://callisto.mitre.org/

[3] http://wordfreak.sourceforge.net/

[4] http://gate.ac.uk/

[5] http://mmax.eml-research.de/

The development of our biomedical annotation tools was driven by two important needs, essential for creating useful text corpora: i) accuracy and consistency of the annotations, and ii) usability of the data. The major aim of this work is therefore two-fold: i) to provide a friendly environment for curators and ii) to take advantage of the multiple informational resources available, enhancing the annotation process as much as possible.

In this sense, the baseline requirements of our tools were interoperability with other tools/modules and flexibility in terms of annotation schemas and data exchange formats. Annotation schemas should be made as general as possible, covering major biomedical classes and thus, enabling (partial) schema interchange. Also, document annotation may comprise both syntactic (POS information) and semantic annotations (BioNER information).

The main aim of the annotation environment presented here is to provide common text processing modules and to enable automatic and manual document annotation. The text processing pipeline was modelled with minimal assumptions on their dependences and application ordering. Tokenisation, sentence splitting and stopword removal are the basic text processing steps, and typically do not rely on previous preprocessing, whereas chunk parsing as well as BioNER may be based on POS annotation. Not only the tools should be able to deal with multilayer annotation, as annotation processes should not have precedence over one another, i.e. semantic annotation may occur after or before POS tagging.

Furthermore, neither automatic nor manual annotation processes are considered mandatory. Typically, manual annotation is time-consuming and should be considered a later step, accounting for false positive matches (term homonymy) and miss annotations (term synonymy and term novelty). However, it is up to the user to decide whether to trigger one or the two processes.

4 Implementation

The implementation of our tools devised the following components/modules:

- an input/output module enabling the conversion of documents for common file formats (such as PDF and HTML) to plain text;
- a pre-processing module embracing XML-based text structuring (the title, authors, journal, abstract and the location of major sections are tagged), tokenisation and stopword removal;
- a default annotation schema embracing all major biological entity classes (genes, proteins, compounds and organisms) and some uncommon, although valuable classes (laboratory techniques and physiological states);
- a lexicon-based biomedical annotator which supports the construction of customised dictionaries as well as user-defined rules and lookup tables;
- an user-friendly annotation viewer based on Cascade Style Sheets (CSS) that allows the user to verify and correct annotations and refine dictionary contents.

Additionally, it is important to note that unlike many previous approaches our tools are able to handle both abstracts and full text documents indistinctively. The latter will undoubtedly give an increasing amount of useful information in most cases.

4.1 Lexical Resources

The tool supports two kinds of lexical resources: lookup tables and dictionaries. The authors have prepared lookup lists of standard laboratory techniques and general physiological states. Also, the user may create general or particular dictionaries from major biomedical databases such as BioCyc[6], UniProt[7] or ChEBI[8] and integrated databases such as Biowarehouse[9] (Figure 1). Each data source is characterised in terms of the embraced biological classes and organism (if it is a multi-organism source). The user may decide to include all contents or select just a few, depending on the purpose of the dictionary.

Database copyrights are preserved as there is no content distribution with the tool. In order to deploy any loader, the user has to download the contents from the corresponding source.

[6] http://biocyc.org/
[7] http://www.uniprot.org/
[8] http://www.ebi.ac.uk/chebi/
[9] http://biowarehouse.ai.sri.com/

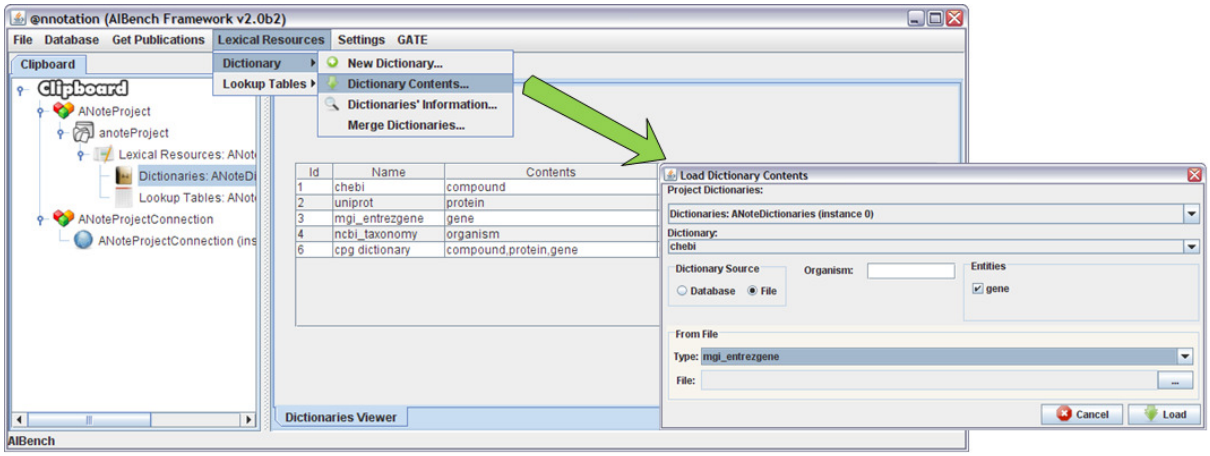

Figure 1. Deploying the construction of a new dictionary using available data loaders.

On the other hand, all created resources are kept in relational format (currently, on MySQL database engine) and thus, allow eventual sharing.

4.2 Annotation Schemas

The default semantic annotation schema was created by the authors and aims at tracking down major biological entities. Currently, the system accounts for a total of 14 biological classes as follows:

- gene
  - metabolic gene
  - regulatory gene
- protein
  - transcription factor
  - enzyme
- pathway
- reaction
- compound
- organism
- DNA
- RNA
- physiological state
- laboratory technique

This schema allows the user to identify molecular entities that may describe different levels of biological organisation and thus, lead to a better insight in functional description of cellular processes.

For instance, a physiological state is frequently characterised by particular level of defined biological entities, like compounds catalysed by certain enzymes, which in turn are encoded by the respective genes. Besides common annotation, this schema also supports annotation linking to lexical resources (Figure 2), i.e., it identifies the dictionary entry that triggered each tagging as well as the normalised term (the "concept label" that gathers together known variants and synonyms of a given term).

The ability to use other annotation schemas is considered a premise of tool interoperability and data re-use. As such, annotation schemas derived from the GENIA ontology (Kim et al., 2003), a formal model of cell signaling reactions in human, or used in challenges such as Biocreative, often referenced by the research community as gold standards, were accounted for. It is possible to choose which schema to use on a given annotation task and also to translate from one schema to another. Additionally, we devise the incorporation of new schemas as long as the user specifies tagging and mapping functions.

Regarding POS, the premise is similar and thus, we chose to incorporate GATE for the development language processing components. GATE provides a reusable design and a set of prefabricated software building blocks (namely tokenizers, sentence splitters and POS taggers) that can be used, extended and customised for specific needs. Also, its component-based model allows for easy coupling and decoupling of the processors, thereby facilitating comparison of alternative configurations or different implementations of the same module (e.g., different parsers). At Figure 2, we illustrate an example of POS tagging output.

Figure 2. Small piece of an annotated document using the default annotation schema and GATE default POS tagging.

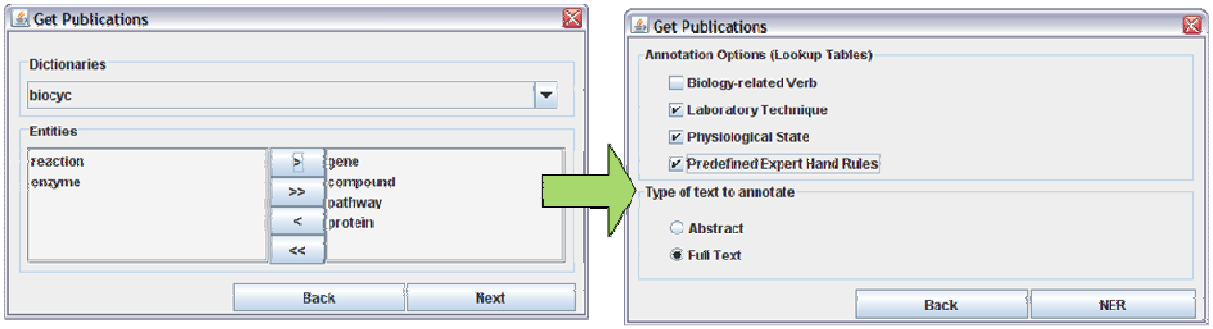

Figure 3. Configuring the automated lexical-based BioNER process.

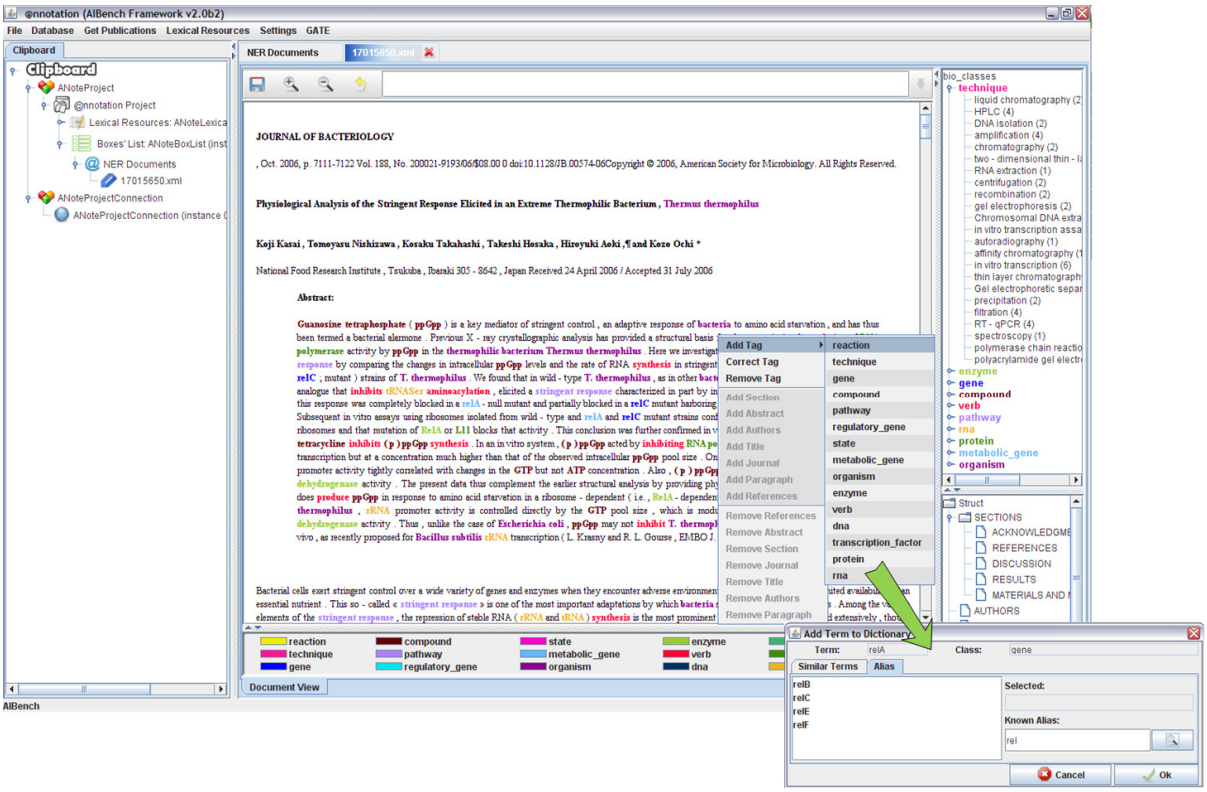

Figure 4. Snapshot of the manual annotation environment.

4.3 Automatic Annotation

The conversion of source formats into plain text is carried out by freeware programs such as Xpdf[10] (Windows or Linux) and pdftotext[11] (Mac OS). The process of XML-oriented document structuring was implemented by the authors using simple pattern matching. Documents (abstracts or full-texts) are submitted to tokenising and stopword removal processes, implemented using Lingua::PT::PLNbase and Lingua::StopWords Perl modules, respectively.

Following the pre-processing step, lexicon-based BioNER is sustained by a specialised rewriting system developed by the authors upon the Text::RewriteRules Perl module. The user specifies the supporting dictionary and the set of biological classes to be annotated (Figure 3). Lookup tables and general templates may also be included. Furthermore, the process can be deployed over abstracts or full-texts.

The system attempts to match terms against dictionary and lookup table contents, checking for different term variants (e.g. hyphen and apostrophe variants) and excluding too short terms (less than 3-character long). Annotation gives preference to longest term matching, tracking up to hepta-grams (i.e. 7-word composition).

Additional patterns account for previously unknown terms and term variants. For example, the template "([a-z]{3}[A-Z]+\d*)" (a sequence of three lower-case letters followed by an upper-case letter and a sequence of zero or more digits) is used to identify candidate gene names while the categorical nouns "ase" and "mRNA" track down possible enzyme and RNA mentions, respectively. Besides class identification, the system also sustains term normalisation, grouping all term variants around a "common name" for visualisation and statistical purposes.

4.4 Manual Annotation

The manual annotation environment accounts for the review of automatic annotations by experts and the enhancement of the lexical resources. Also, manually curated documents are intended to be further used as training corpora to build annotation, classification or other generalised learning models regarding biomedical contents.

Although the actual corpus file with annotation is encoded in XML, the annotators work on

[10] http://www.foolabs.com/xpdf/
[11] http://www.bluem.net/downloads/pdftotext_en/

a CSS-styled view which is much more user-friendly (Figure 4). Furthermore, a query view is used to depict the relation of the annotated terms with dictionary entries.

When the user revises dictionary-based annotation and corrects or adds annotations, the dictionary is updated with such previously unknown or mischaracterised information. Therefore, this process has two major outputs: high-quality annotation and dictionary enrichment. The latter is a classical example of a process of learning by experience that accounts for well-known biological issues such as term novelty, term synonymy and term homonymy. Term novelty and the association of synonyms are far from being adequately tackled as they will depend on expert's knowledge, which is limited and often outdated just like dictionaries. However, the disambiguation of distinct mentions using the same term (e.g. same gene, protein and RNA name) is a classical example where manual curation is invaluable.

Also, users may cooperate on curation tasks, sharing locally processed documents and taking advantage of dictionaries that have been refined by other users.

5 Conclusions

The need for user-friendly and interoperable semantic annotation tools is indisputable in BioTM. Research benefits greatly from the reuse of data (such as annotated corpora) and the capacity to interchange tools (namely POS and semantic taggers). However, this is only possible if tools are devised for this purpose, i.e., if they account for general annotation as well as annotation interchange and if processing tools are prepared to account for distinct annotation schemas. On the other hand, annotation is a laborious and time-consuming task that requires from the curators both expertise on the subjects and critical judgment. In this sense, it is very important that annotation tools take advantage of data mining models and available knowledge resources, minimising manual curation efforts, and at the same time, provide for a user-friendly environment.

In this work, a contribution to these issues is provided, with the development of a novel interoperable and user-friendly software tool for biomedical annotation. Its primary contributions are as follows: the ability to process abstract and full-texts interchangeably; a basic semantic annotation schema encompassing embracing all major biomedical entity classes (genes, proteins, compounds and organisms) and some uncommon, although valuable classes (laboratory techniques and physiological states); the ability to use standard annotation schemas such as GENIA; a preprocessing module capable of converting documents from common file formats (such as PDF and HTML) to plain text and then, tokenise and remove stopword from such texts; a lexicon-based biomedical annotator for annotating biomedical texts which allows the construction of customised dictionaries as well as user-defined rules and lookup tables; a user-friendly annotation view that allows the user to verify and correct annotations and refine dictionary contents.

The tool can be used as a stand-alone environment or it can be integrated in a more comprehensive BioTM framework. Currently, it is incorporated in the @Note Biomedical Text Mining workbench[12] (Lourenço et al., 2008). Here, tool interoperability enables automatic information retrieval (PubMed keyword-based query and document retrieval from open-access and subscribed web-accessible journals) as well as mining experiments (using annotated corpora to construct BioNER models).

Future work includes the enhancement of annotation skills based on curator suggestions and the implementation of several measures to minimize discrepancies of inter-annotation and maintain the quality of annotation. Semantic type checking and detection of anomalies in the resulting annotations are devised as the first steps.

The tools are freely available from http://sysbio.di.uminho.pt/anote.php.

Acknowledgments

This work is partly funded by the research projects recSysBio (ref. POCI/BIO/60139/2004) and MOBioPro (ref. POSC/EIA/59899/2004) financed by the Portuguese Fundação para a Ciência e Tecnologia. The work of Sónia Carneiro is supported by a PhD grant from the Fundação para a Ciência e Tecnologia (ref. SFRH/BD/22863/2005).

References

S. Ananiadou, D. B. Kell and J. I. Tsujii (2006). Text mining and its potential applications in systems biology. *Trends Biotechnol., 24,* 571-579.

[12] http://sysbio.di.uminho.pt/anote.php

H. Cunningham, D. Maynard, K. Bontcheva and V. Tablan (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02).

R. A. A. Erhardt, R. Schneider and C. Blaschke (2006). Status of text-mining techniques applied to biomedical text. *Drug Discovery Today, 11,* 315-325.

K. Fundel and R. Zimmer (2006). Gene and protein nomenclature in public databases. *BMC Bioinformatics, 7*.

D. Hanisch, K. Fundel, H. T. Mevissen, R. Zimmer and J. Fluck (2005). ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics, 6*.

Z. Z. Hu, M. Narayanaswamy, K. E. Ravikumar, K. Vijay-Shanker and C. H. Wu (2005). Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics, 21,* 2759-2765.

L. J. Jensen, J. Saric and P. Bork (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics, 7,* 119-129.

J. D. Kim, T. Ohta, Y. Tateisi and J. Tsujii (2003). GENIA corpus--semantically annotated corpus for bio-textmining. *Bioinformatics, 19 Suppl 1,* i180-i182.

J. D. Kim, T. Ohta and J. Tsujii (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics, 9*.

Z. Kou, W. W. Cohen and R. F. Murphy (2005). High-recall protein entity recognition using a dictionary. *Bioinformatics, 21 Suppl 1,* i266-i273.

Kulick S, Bies A, Liberman M, Mandel M, McDonald R, Palmer M, Schein A and Ungar L (2004). Integrated Annotation for Biomedical Information Extraction. NAACL/HLT Workshop on Linking Biological Literature,Ontologies and Databases: Tools for Users (pp. 61-68).

A. Lourenço, R. Carreira, S. Carneiro, P. Maia, D. Glez-Peña, F. Fdez-Riverola, E. C. Ferreira, I. Rocha and M. Rocha (2008). @Note: a flexible and extensible workbench for Biomedical Text Mining. *Submitted to BMC Bioinformatics*.

T. Morton and J. LaCivita (2003). WordFreak: an open tool for linguistic annotation. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations (pp. 17-18). NJ, USA: Association for Computational Linguistics Morristown.

S. Mukherjea, L. V. Subramaniam, G. Chanda, S. Sankararaman, R. Kothari, V. Batra, D. Bhardwaj and B. Srivastava (2004). Enhancing a biomedical information extraction system with dictionary mining and context disambiguation. *Ibm Journal of Research and Development, 48,* 693-701.

H. M. Muller, E. E. Kenny and P. W. Sternberg (2004). Textpresso: An ontology-based information retrieval and extraction system for biological literature. *Plos Biology, 2,* 1984-1998.

J. Natarajan, D. Berrar, C. J. Hack and W. Dublitzky (2005). Knowledge discovery in biology and biotechnology texts: A review of techniques, evaluation strategies, and applications. *Critical Reviews in Biotechnology, 25,* 31-52.

N. Okazaki and S. Ananiadou (2006). Building an abbreviation dictionary using a term recognition approach. *Bioinformatics, 22,* 3089-3095.

L. Shi and F. Campagne (2005). Building a protein name dictionary from full text: a machine learning term extraction approach. *BMC Bioinformatics, 6,* 88.

C. J. Sun, Y. Guan, X. L. Wang and L. Lin (2006). Biomedical named entities recognition using conditional random fields model. *Fuzzy Systems and Knowledge Discovery, Proceedings, 4223,* 1279-1288.

L. Tanabe, N. Xie, L. H. Thom, W. Matten and W. J. Wilbur (2005). GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics, 6*.

L. Yeganova, L. Smith and W. J. Wilbur (2004). Identification of related gene/protein names based on an HMM of name variations. *Computational Biology and Chemistry, 28,* 97-107.

P. Zweigenbaum, D. mner-Fushman, H. Yu and K. B. Cohen (2007). Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics, 8,* 358-375.