

Chapter XVII

A Software Tool for Biomedical Information Extraction (And Beyond)

Burr Settles

University of Wisconsin-Madison, USA

ABSTRACT

ABNER (A Biomedical Named Entity Recognizer) is an open-source software tool for text mining in the molecular biology literature. It processes unstructured biomedical documents in order to discover and annotate mentions of genes, proteins, cell types, and other entities of interest. This task, known as named entity recognition (NER), is an important first step for many larger information management goals in biomedicine, namely extraction of biochemical relationships, document classification, information retrieval, and the like. To accomplish this task, ABNER uses state-of-the-art machine learning models for sequence labeling called conditional random fields (CRFs). The software distribution comes bundled with two models that are pre-trained on standard evaluation corpora. ABNER can run as a stand-alone application with a graphical user interface, or be accessed as a Java API allowing it to be re-trained with new labeled corpora and incorporated into other, higher-level applications. This chapter describes the software and its features, presents an overview of the underlying technology, and provides a discussion of some of the more advanced natural language processing systems for which ABNER has been used as a component. ABNER is open-source and freely available from <http://pages.cs.wisc.edu/~bsettles/abner/>

INTRODUCTION

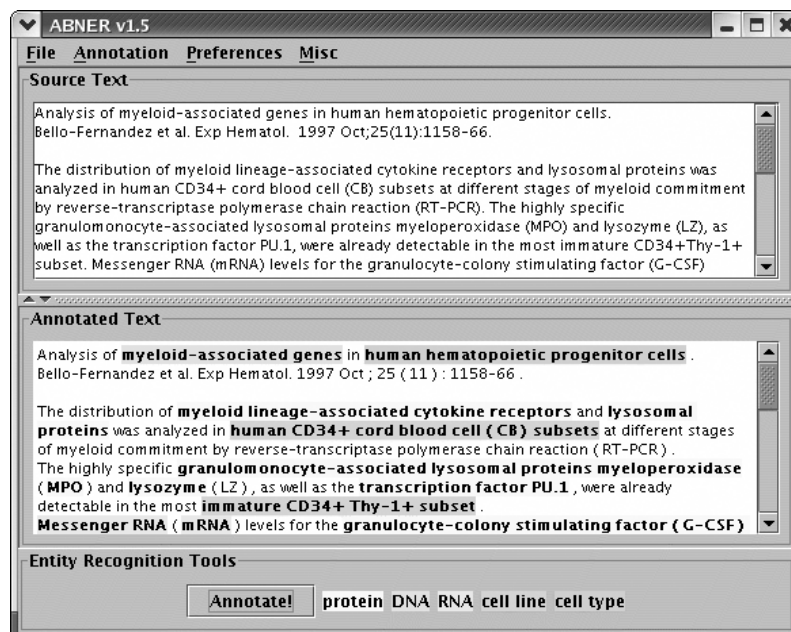
Efforts to organize the wealth of biomedical knowledge in the primary literature have resulted in hundreds of databases and other resources (Bateman, 2008), providing scientists with access to structured biological information. However, with nearly half a million new research articles added to PubMed annually (Soteriades & Falagas, 2005), the sheer volume of publications and complexity of the knowledge to be extracted is beyond the means of most manual database curation efforts. As a result, many of these resources struggle to remain current. Automated *information extraction* (IE), or at least automated assistance for such extraction tasks, seems a natural way to overcome these information management bottlenecks.

Named entity recognition (NER) is a subtask of IE, focused on finding mentions of various *entities* that belong to semantic classes of interest. In the biomedical domain, entities of interest are usually references to genes, proteins, cell

types, and the like. Accurate NER systems are an important first step for many larger information management goals, such as automatic extraction of biologically relevant relationships (e.g., protein-protein interactions or sub-cellular location of gene products), biomedical document classification and retrieval, and ultimately the automatic maintenance of biomedical databases.

In order to facilitate and encourage research in the area of biomedical NER, several “bake-off” style competitions have been organized, in particular the NLPBA shared task (Kim et al., 2004) and the BioCreative challenge (Yeh et al., 2005). For these events, several research teams rapidly design, build, and submit results for machine learning systems using benchmark annotated text collections. The challenges showcase a variety of approaches to the problem, and provide a wealth of insights into what sorts of models and features are most effective. However, few of the resulting systems have been made publicly available for researchers working in related areas of natural language processing (NLP) in biomedicine.

Figure 1. A screenshot of ABNER’s graphical user interface



8 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/software-tool-biomedical-information-extraction/23068

Related Content

Using a Genetic Algorithm and Markov Clustering on Protein–Protein Interaction Graphs

Charalampos Moschopoulos, Grigorios Beligiannis, Spiridon Likothanassis and Sophia Kossida (2012). *International Journal of Systems Biology and Biomedical Technologies* (pp. 35-47).

www.irma-international.org/article/using-genetic-algorithm-markov-clustering/67105

In Silico Biology: Making the Most of Parallel Computing

Dimitri Perrin, Heather J. Ruskin and Martin Crane (2010). *Biocomputation and Biomedical Informatics: Case Studies and Applications* (pp. 55-74).

www.irma-international.org/chapter/silico-biology-making-most-parallel/39603

Extracting Patient Case Profiles with Domain-Specific Semantic Categories

Yitao Zhang and Jon Patrick (2009). *Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration* (pp. 273-287).

www.irma-international.org/chapter/extracting-patient-case-profiles-domain/23065

RETRACTED: Political Discourse as Sliding Mode Manifestations

Ekaterina Yuryevna Aleshina (2022). *International Journal of Applied Research in Bioinformatics* (pp. 1-10).

www.irma-international.org/article/retracted-political-discourse-as-sliding-mode-manifestations/282690

An Optimized Semi-Supervised Learning Approach for High Dimensional Datasets

Nesma Settouti, Mostafa El Habib Daho, Mohammed El Amine Bechar and Mohammed Amine Chikh (2018). *Applying Big Data Analytics in Bioinformatics and Medicine* (pp. 294-321).

www.irma-international.org/chapter/an-optimized-semi-supervised-learning-approach-for-high-dimensional-datasets/182952