Chapter 3.17
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/

DOI: 10.4018/978-1-60566-274-9.ch017

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.

I first released ABNER (Settles, 2005) in July 2004 as a demonstrational graphical user interface (GUI) for the system I developed as part of the NLPBA shared task challenge (Settles, 2004). In March 2005, a revised, open-source version of the software was released with some performance improvements and a new Java application programming interface (API). The goal is to encourage others to write custom interfaces to the core NER software, allowing it to be integrated into other, more sophisticated biomedical information management systems. ABNER also supports training new models on corpora labeled for different knowledge domains (e.g., particular organisms, since gene naming conventions vary from species to species).

Figure 1 shows a screenshot of the intuitive GUI when ABNER is run as a stand-alone application. Text can be typed in manually or loaded from a file (top window), and then automatically tagged for multiple entities in real time (bottom window). Each entity type is highlighted with a unique color for easy visual reference, and tagged documents can be saved in a variety of annotated file formats. The application also has options for processing plain text documents on the file system in batch mode offline.

ABNER has built-in functionality for tokenization and sentence segmentation, which are fairly robust to line breaks and biomedical abbreviations (users can choose to bypass these features in favor of their own text preprocessing as well). The bundled ABNER application is implemented in Java and is therefore platform-independent, and has been tested on Linux, Solaris, Mac OS X, and Windows.

The basic ABNER distribution includes two built-in entity-tagging models trained on the NLPBA (Kim et al., 2004) and BioCreative (Yeh et al., 2005) corpora. The first is a modified version of the GENIA corpus (Kim et al., 2003), contain-