BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 14 2005, pages 3191–3192
doi:10.1093/bioinformatics/bti475

Data and text mining

ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text

Burr Settles

Department of Computer Sciences and Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 52706, USA

Received on April 1, 2005; revised on April 21, 2005; accepted on April 26, 2005
Advance Access publication April 28, 2005

ABSTRACT

Summary: ABNER (A Biomedical Named Entity Recognizer) is an open source software tool for molecular biology text mining. At its core is a machine learning system using conditional random fields with a variety of orthographic and contextual features. The latest version is 1.5, which has an intuitive graphical interface and includes two modules for tagging entities (e.g. protein and cell line) trained on standard corpora, for which performance is roughly state of the art. It also includes a Java application programming interface allowing users to incorporate ABNER into their own systems and train models on new corpora.

Availability: ABNER is available as an executable Java archive and source code from http://www.cs.wisc.edu/~bsettles/abner/

Contact: bsettles@cs.wisc.edu

1 INTRODUCTION

Interest in developing effective tools for natural language processing (NLP) tasks in biomedical literature has been increasing in recent years. The tasks offer scientific challenges—established NLP techniques do not port easily to the biomedical domain—but there is also a practical need to effectively curate, organize and retrieve information automatically from textual sources. Named entity recognition, the NLP task of identifying words and phrases belonging to certain classes (e.g. protein and cell line), is an important first step for many larger information management goals. The current state of the art yields F_1 scores with *exact* boundary matching around 70 (Kim *et al*., 2004; Yeh *et al*., 2004), but few systems with published results in this range are freely available.

ABNER (A Biomedical Named Entity Recognizer) version 1.0 was released in July 2004 as a free, user-friendly interface to a high-performing system developed for the NLPBA 2004 Shared Task (Settles, 2004). Version 1.5 was released open source in March 2005 with some performance improvements and a customizable application programming interface (API).

2 SOFTWARE FEATURES

ABNER has an intuitive graphical user interface where text can be typed in manually or loaded from a file and automatically tagged for multiple named entities in real time. A screen shot of the interface is shown in Figure 1. Each entity is highlighted with a unique color (yellow = protein, green = DNA, etc.) for easy visual reference, and tagged documents can be saved in a variety of file formats. The software can also annotate plain text files in batch mode. Users can pre-tokenize input text, or make use of ABNER's built-in tokenization, which is quite robust to wrapped lines and biomedical abbreviations. The bundled ABNER application is platform-independent and has been tested on Linux, Windows XP, Solaris and Mac OSX. The distribution includes two built-in entity tagging modules that are trained and evaluated on the standard NLPBA (Kim *et al*., 2004) and BioCreative[1] (Yeh *et al*., 2004) corpora. Performance details for both modules are presented in Section 4.

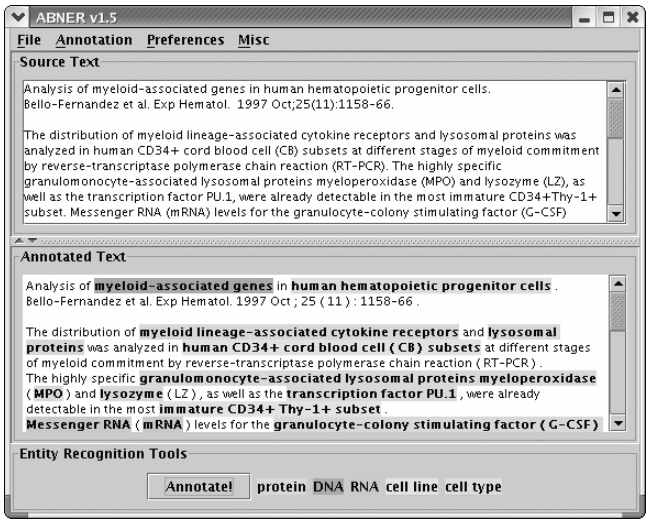

Fig. 1. A screen shot of ABNER's graphical user interface.

The Java API allows users to write custom interfaces to ABNER modules or incorporate them into larger biomedical NLP systems. The API also includes routines for training new modules on other corpora. (This may be necessary for tasks that are organism-specific or require tagging conventions not reflected by the built-in modules.) The source code is also available under the terms of the Common Public License.

[1]This was previously distributed as part of a command-line tool called YAGI (Yet Another Gene Identifier), which has been deprecated.

3 ALGORITHMS AND IMPLEMENTATION

Conditional random fields (CRFs) are undirected statistical graphical models, a special case of which corresponds to conditionally trained finite-state machines well suited for labeling and segmenting sequence data (Lafferty *et al*., 2001). Named entity recognition can be framed as a sequence labeling problem: words in a sentence are tokens to be assigned labels by states in the CRF framework.

Let $\mathbf{o} = \langle o_1, o_2, \ldots, o_n \rangle$ be a sequence of observed words of length n. Let L be a set of labels (protein, DNA, other, etc.) corresponding to states in a finite-state machine. Then $\mathbf{l} = \langle l_1, l_2, \ldots, l_n \rangle$ is a sequence of labels from L assigned to words in the input sequence $\mathbf{o}$. A first-order linear-chain CRF defines the conditional probability of a label sequence given an input sequence to be:

$$P(\mathbf{l}|\mathbf{o}) = \frac{1}{Z_\mathbf{o}} \exp\left(\sum_{i=1}^{n}\sum_{j=1}^{k} \lambda_j f_j(l_{i-1}, l_i, \mathbf{o}, i)\right),$$

where $Z_\mathbf{o}$ is a normalization factor over all possible label sequences, f_j is one of the k binary functions describing a feature at position i in sequence $\mathbf{o}$ and λ_j is a weight for that feature. For example, given the text '... the ATPase...' f_j might be the feature WORD=*ATPase* and have value 1 along the transition where l_{i-1} is the label state other ('the' is a non-entity) and l_i is the label state protein. Other features with value 1 along this transition are CAPITALIZED, MIXED-CASE and SUFFIX=*ase*. The learned weight λ_j should be positive for a feature correlated with the target label, negative for a feature that is anti-correlated and near zero for a relatively uninformative feature. The weights are set to maximize the conditional log-likelihood of m labeled sequences in a training set $D = \{\langle \mathbf{o}, \mathbf{l} \rangle_{(1)}, \ldots, \langle \mathbf{o}, \mathbf{l} \rangle_{(m)}\}$:

$$LL(D) = \sum_{i=1}^{m} \log\left(P(\mathbf{l}_{(i)}|\mathbf{o}_{(i)})\right) - \sum_{j=1}^{k} \frac{\lambda_j^2}{2\sigma^2},$$

where the second sum is a Gaussian prior over feature weights to help to prevent overfitting due to sparsity in D. If training sequences are fully labeled, $LL(D)$ is convex and the model is guaranteed to converge optimally. New sequences can then be labeled with the Viterbi algorithm. For more details, see Lafferty *et al*. (2001).

ABNER's default feature set comprises orthographic and contextual features, mostly based on regular expressions and neighboring tokens. The feature set is slightly modified from previous work (Settles, 2004) for improved performance, and can be viewed/modified in the source code distribution. Note that ABNER currently does not use syntactic or semantic features. Research indicates that such features can improve performance slightly, but presently they are not dynamically generated by ABNER.

The system is written entirely in Java using graphical window objects from the Swing library. The CRF models are implemented with the MALLET toolkit (http://mallet.cs.umass.edu/), which uses a quasi-Newton method called L-BFGS (Nocedal and Wright, 1999) to find the optimal feature weights efficiently. Tokenization is performed by a deterministic finite-state scanner built with the JLex tool (http://www.cs.princeton.edu/~appel/modern/java/JLex/).

4 EVALUATION

The NLPBA corpus is a modified version of the GENIA corpus (Kim *et al*., 2003), containing five entities labeled for 18 546 training sentences and 3856 evaluation sentences. The BioCreative corpus contains only one entity subsuming genes and gene products (proteins, RNA, etc.) labeled for 7500 training sentences and 2500 evaluation sentences. ABNER tagged the NLPBA corpus at a rate of 864 words (33 sentences) per second, and the BioCreative corpus at a rate of 1260 words (48 sentences) per second on a 500 MHz Pentium III running Linux with 512 MB memory (speeds will vary among different tagging modules and machines).

Table 1. Evaluation of ABNER's two tagging modules

Corpus	Entity	R	P	F_1	$(S-F_1)$
NLPBA	Overall	72.0	69.1	70.5	(82.0)
	protein	77.8	68.1	72.6	(84.9)
	DNA	63.1	67.2	65.1	(76.1)
	RNA	61.9	61.3	61.6	(78.5)
	cell line	58.2	53.9	56.0	(68.2)
	cell type	65.6	79.8	72.0	(82.1)
BioCreative	protein/gene	65.9	74.5	69.9	(83.7)

Recall, precision and F_1 reflect exact boundary matching. $S - F_1$ is a soft F_1 score where either the left- or right-boundary must be correct, but a one-token error on the other boundary is tolerated.

Table 1 presents evaluation results in terms of recall [TP/(TP + FN)], precision [TP/(TP + FP)], and F_1 score [$(2 \times R \times P/(R + P)$], where TP means true positives, FN means false negatives and FP means false positives. To the author's knowledge, these figures are competitive with the best published results on these corpora at this time. It is important to note that the quality of biomedical NLP systems can vary by organism (Hirschman *et al*., 2004), thus training ABNER with novel, organism-specific corpora (with a potentially augmented feature set) may be advisable for some applications.

ACKNOWLEDGEMENTS

Thanks to Mark Craven for his support of this project. Research related to development of this software supported by NLM grant 5T15LM007359 and NIH grant R01 LM07050-01.

REFERENCES

Hirschman,L., Colosimo,M., Morgan,A., Colombe,J. and Yeh,A. (2004) Task 1B: gene list task. In *Proceedings of the Critical Assessment of Information Extraction Systems in Biology* (*BioCreAtIvE*) *Workshop*, Grenada, Spain.

Kim,J. *et al*. (2003) GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, **19**(Suppl. 1), i180–i182.

Kim,J., Ohta,T., Tsuruoka,Y., Tateisi,Y. and Collier,N. (2004) Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications* (*NLPBA*), Geneva, Switzerland, pp. 70–75.

Lafferty,J., McCallum,A. and Pereira,F. (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. San Francisco, CA. In *Proceedings of the International Conference on Machine Learning*, Williamstown, MA. Morgan Kaufmann, pp. 282–289.

Nocedal,J. and Wright,S.J. (1999) *Numerical Optimization*. Springer, New York, NY, pp. 224–233.

Settles,B. (2004) Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications* (*NLPBA*), Geneva, Switzerland, pp. 104–107.

Yeh,A., Hirschman,L., Morgan,A. and Colosimo,M. (2004) Task 1A: gene-related name mention finding evaluation. In *Proceedings of the Critical Assessment of Information Extraction Systems in Biology* (*BioCreAtIvE*) *Workshop*, Grenada, Spain.