

## GeneTaggerCRF: An Entity Tagger for Recognizing Gene Names in Text

Rishi Talreja<sup>\*†1</sup> Andrew I. Schein<sup>1</sup> R. Scott Winters<sup>2</sup> Lyle H. Ungar<sup>1</sup>

Departments of <sup>1</sup>Computer and Information Science and <sup>2</sup>Pediatrics, University of Pennsylvania, 3330 Walnut Street, Philadelphia, PA 19104 USA

<sup>2</sup>The Children's Hospital of Philadelphia, 34th and Civic Center Blvd. Philadelphia PA 19104 USA

### ABSTRACT

**Summary:** GeneTaggerCRF is an application for tagging gene names and references in text. It is based on the MALLET natural language processing package and uses a machine learning technique called conditional random fields. It is the first installment of a set of information extraction tools called BioSFIER (Biology Software For Information Extraction and Retrieval). We tested GeneTaggerCRF's performance by 10-fold cross validation on 190 MEDLINE abstracts pertaining to Neuroblastoma cancer genetics. Performance was 0.93 precision and 0.60 recall.

**Availability:** The software is available at [http://www.cis.upenn.edu/datamining/software\\_dist/biosfier](http://www.cis.upenn.edu/datamining/software_dist/biosfier)

**Contact:** [rtalreja@uiuc.edu](mailto:rtalreja@uiuc.edu)

### INTRODUCTION

Recent technological advancements have resulted in the explosive growth of both biomedical data and literature. Automatic identification of gene mentions in text by named entity taggers can greatly facilitate research in biomedical disciplines. Named entity taggers add additional query mechanisms for information retrieval applications. An immediate application for such entity taggers is in indexing documents by entities they contain. These indices facilitate advanced query interfaces, for instance, querying documents by the genes they mention, in addition to recognition of similarities between documents based on the entities referenced. Alternatively, named entity taggers are viewed as an important component of an information extraction system: a system that automatically extracts facts of interest from text documents, *e.g.* see (Donaldson *et al.*, 2003). We present GeneTaggerCRF, a named entity tagger for genes that performs at 0.93 precision and 0.60 recall according to cross-validation on a training set of MEDLINE abstracts.

A number of algorithms have been used for identifying gene name instances in text, including rule-based systems (Tanabe and Wilbur, 2002; Narayanaswamy *et al.*, 2003), hidden Markov models (Collier *et al.*, 2000), and support vector machines (Kazama *et al.*, 2002). GeneTaggerCRF builds upon previous efforts at gene entity tagging in several respects. The

main advancement GeneTaggerCRF makes in entity tagging is that it is based on a new machine learning technique for labeling sequence data called conditional random fields (Lafferty *et al.*, 2001). Second, our definition of gene (described below) used to produce training material differs from previous work and should prove general to many applications. Finally, the underlying data on which the method is trained including definitions and annotation guidelines is made available, facilitating reproduction of our results and improvement by the community at large.

GeneTaggerCRF is the first of a number of tools under construction as part of an information extraction toolkit called BioSFIER (Biological Software For Information Extraction and Retrieval). A goal of our on-going project is to produce various forms of syntactic and semantic annotation of biomedical text documents (Kulick *et al.*, 2003) to aid in information extraction. GeneTaggerCRF is available for download and will be retrained as we acquire more annotated training data. The training data, including detailed finalized annotation guidelines will be made available in a forthcoming publication. Data used in these experiments are available by contacting the authors.

### DATA, TASK, AND GENE DEFINITION

Our task was to train a gene recognizer in order to find all genes in given text, where a gene is defined as in (UPenn Biomedical Information Extraction Group, 2003). As a brief overview, we define a gene as being a conceptual entity distinct from specific instances including raw sequence information, positions on a chromosome, or specific transcripts or subunits, all of which may be captured under separate entity categories or as relationships between entity categories. Further, the gene definition is head of a hierarchy consisting of more descriptive subtypes (*i.e.* *genomic* and *protein product*) as well as an underspecified class (*i.e.* *gene generic*). In our present work we blur the distinctions between the various gene products and the gene concept into a single *gene* entity for making predictions. The definition of gene our algorithm is trained to recognize breaks the precedent set in an earlier published annotation project, (GENIA, 2004), by making genes and their physical representations separate entities. While the distinction may not be germane for the purpose the GENIA corpus serves, it is sensible when discussing

<sup>\*</sup>To whom correspondence should be addressed.

<sup>†</sup>Present address: Department of Computer Science, University of Illinois, 1304 W. Springfield Ave., Urbana, IL 61801 USA

genes and their mutations, which is the ultimate concern of our project.

As an illustration, suppose we are given the following sentence:

*Cytogenetic abnormalities and their lack of relationship to the Asp816Val c-kit mutation in the pathogenesis of mastocytosis.*

Then the word c-kit would be tagged as a gene. On the other hand, in the example below:

*two single-point mutations of codon 13 were shown.*

the phrase codon 13 is not annotated as a gene since it refers to a set of positions within the genomic sequences of a gene, but is not itself a gene.

## ALGORITHM

The function named entity taggers perform is to take an observed sequence of word tokens ( $O_p$ , e.g. the word sequence in a MEDLINE abstract) and output a label sequence of tags ( $T_p$ ) where tag variables  $T_p$  take on one of three categories  $\{b, c, o\}$  encoding: begin gene, continue gene, and outside of (i.e. not a) gene. The three categories segment the text into gene segments and non-gene segments. In the past, Hidden Markov models (HMMs) have been used to perform similar biomedical entity labeling tasks (Collier et al., 2000), and stochastic grammars have been applied to entity tagging in other domains. These models represent joint probability to word-token and label sequences ( $\Pr(T, P)$ ) through a random process where the labels stochastically generate the word tokens ( $\Pr(T, O) = \Pr(T) \Pr(O|T)$ ). The probability of a tag sequence given the observed token sequence ( $\Pr(T|O)$ ) is computed through application of Bayes rule.

Such generative models require making false independence assumptions when multiple lexical features of the individual tokens are used as predictors of the corresponding tags. In contrast, conditional probability models directly assign a probability to the labels given the word-tokens ( $\Pr(T|O)$ ) without application of Bayes rule, and do not require any false independence assumptions. Conditional models are attractive when many predictors are used in tagging. Conditional random field (CRF) (Lafferty et al., 2001) models are one such conditional probability method for sequence tagging. To build a gene tagger we employed the MALLETT (McCallum, 2002) implementation of the CRF model:

$$\Pr(T|O) \propto \exp\left(\sum_p \sum_i \lambda_i \cdot f_i(O_p, T_{p-1}, T_p)\right). \quad (1)$$

Evaluation on test data is performed through the use of Viterbi decoding which finds the maximum likelihood sequence of tags on the observed data.

The functions  $f_i$  are predictive features of the data that are weighted by  $\lambda_i$  terms according to the maximum likelihood method of parameter tuning. We used a variety of binary-valued predictive feature functions (see Table 1 for examples)

Feature	Example
CAPITALIZED	Gene
ALLCAPS	CYP
MIXEDCAPS	BetaY
CONTAINSDIGITS	CYP450
ALLDIGITS	45
NUMERICAL	15,000
ALPHANUMERIC	a45b
CONTAINSDASH	c-kit
LONELYINITIAL	P
SINGLECHAR	c
PUNC	,

Table 1: Some binary features used in GeneTaggerCRF along with examples.

in our method in addition to the identity of words immediately in front of and behind the word to be tagged. A complete listing of features employed in training can be found in the distribution documentation for GeneTaggerCRF.

## RESULTS

The performance of GeneTaggerCRF was calculated using:

$$\text{Precision} = \frac{\# \text{ of genes predicted correctly}}{\# \text{ of genes predicted}}$$

$$\text{Recall} = \frac{\# \text{ of genes predicted correctly}}{\# \text{ of genes in text}}$$

We employed the very stringent scoring guideline that both boundaries of a predicted gene must match the expert human annotation in order for a prediction to be deemed correct.

Using these evaluation metrics, we performed 10-fold cross-validation on 190 manually annotated abstracts from the MEDLINE database. GeneTaggerCRF succeeded in tagging the abstracts with 0.93 precision and 0.60 recall. Eventually, we expect to have 2,000 to 10,000 annotated MEDLINE abstracts available for training our tagger, which should significantly improve performance.

## ACKNOWLEDGMENTS

To be included upon acceptance.

## REFERENCES

- Collier, N., Nobata, C., and Tsujii, J. (2000) Extracting the names of genes and gene products with a hidden Markov model. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000)*, Saarbrücken, Germany.

- Donaldson, I., et al. (2003) PreBIND and Textomy—Mining the Biomedical Literature for Protein-Protein Interactions Using a Support Vector Machine. *BMC Bioinformatics*, **4:11**, 1–13.
- GENIA (2004) GENIA Corpus. [Http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/](http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/).
- Kazama, J., Makino, T., Ohta, Y., and Tsujii, J. (2002) Tuning Support Vector Machines for Biomedical Named Entity Recognition. In *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain*, pages 1–8. 2002 Association for Computational Linguistics Annual Conference.
- Kulick, S., Liberman, M., Palmer, M., and Schein, A. (2003) Shallow Semantic Annotation of Biomedical Corpora for Information Extraction. In *Proc. Third Meeting of the Special Interest Group on Text Mining at ISMB 2003*.
- Lafferty, J., McCallum, A., and Pereira, F. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA.
- McCallum, A. K. (2002) MALLETT: A Machine Learning for Language Toolkit. [Http://www.cs.umass.edu/~mccallum/mallet](http://www.cs.umass.edu/~mccallum/mallet).
- Narayanaswamy, M., Ravikumar, K. E., and Vijay-Shanker, K. (2003) A Biological Named Entity Recognizer. In *Pacific Symposium on Biocomputing*, volume 8.
- Tanabe, L. and Wilbur, W. (2002) Tagging gene and protein names in biomedical text. *Bioinformatics*, **18**, 1124–1132.
- UPenn Biomedical Information Extraction Group (2003) BioEntities: Entity Definitions for Oncology Effort. <http://www.cis.upenn.edu/~mamandel/annotators/onco/definitions.html>.