BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 17 2004, pages 3249–3251
doi:10.1093/bioinformatics/bth350

An entity tagger for recognizing acquired genomic variations in cancer literature

Ryan T. McDonald[1,*,†], R. Scott Winters[3,†], Mark Mandel[4], Yang Jin[2], Peter S. White[2,3] and Fernando Pereira[1]

[1]Department of Computer and Information Science, University of Pennsylvania, 3330 Walnut Street, Philadelphia, PA 19104, USA, [2]Department of Pediatrics, University of Pennsylvania, Philadelphia, PA 19104, USA, [3]The Children's Hospital of Philadelphia, 34th and Civic Center Blvd, Philadelphia, PA 19104, USA and [4]Linguistic Data Consortium, University of Pennsylvania, 3401 Walnut St Suite 400A, Philadelphia, PA 19104, USA

Received on April 30, 2004; revised on May 21, 2004; accepted on May 25, 2004
Advance Access publication June 4, 2004

ABSTRACT

Summary: VTag is an application for identifying the type, genomic location and genomic state-change of acquired genomic aberrations described in text. The application uses a machine learning technique called conditional random fields. VTag was tested with 345 training and 200 evaluation documents pertaining to cancer genetics. Our experiments resulted in 0.8541 precision, 0.7870 recall and 0.8192 F-measure on the evaluation set.

Availability: The software is available at http://www.cis.upenn.edu/group/datamining/software_dist/biosfier/.

Contact: ryantm@cis.upenn.edu

INTRODUCTION

The proliferation of biomedical text makes it increasingly difficult for the researchers to track and utilize information relevant to their interests. Automated information extraction techniques can assist in the acquisition, management and curation of these data. A necessary first step is the ability to automatically recognize biomedical entities in text, which is also known in the natural language processing community as named entity recognition.

Development of named entity taggers for biomedical literature has progressed rapidly in recent years. For example, a number of algorithms currently exist for identifying gene name instances in text (Collier *et al*., 2000; Tanabe and Wilbur, 2002; Yu *et al*., 2002; GENIA, 2004, http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/). Another, more complex, entity recognition task, is the identification of genomic variation text mentions, which is applicable both to researchers interested in finding disease–genome associations and to mutation database curators. We present here VTag, a named entity tagger based upon a conditional random field model that addresses the open problem of recognizing variation events in text. To the best of our knowledge, VTag is the first directed effort at automated literature extraction of acquired molecular-genetic variation events, such as point mutations, translocations and deletions. We applied VTag to text describing cancer-associated genomic variation.

TASK

Our task was to develop an automated method that would accurately recognize each component of an acquired genomic aberration (hereafter referred to as a variation event) within a cancer-specific text (UPenn Biomedical Information Extraction Group, 2003, http://www.cis.upenn.edu/∼mamandel/annotators/onco/definitions.html). Briefly, we define a variation event as a specific, one-time alteration at the genomic level, and described at the nucleic acid level, amino acid level or both. Each variation event is identified by the relationship among three variation components: variation type, variation location and variation state (both initial and subsequent states). As an illustration:

> 'All cases with K-ras codon 12 mutations were found to be G to T transversion' (Wang *et al*., 2002).

In this sentence variation component tags would be assigned as follows: transversion, variation type; codon 12, variation location; G, variation state (initial); and T, variation state (subsequent). The relationship among these components defines a single variation event. This entity definition is suitable for a variety of applications (e.g. other genetic diseases) and readily modified to include naturally occurring variations (e.g. single nucleotide polymorphisms). Furthermore, our experience indicates that this definition is generic and capable of capturing the details of diverse variation events

*To whom correspondence should be addressed.

†Both these authors contributed equally to this work.

(e.g. point mutation, translocation, aneuploidy and loss of heterozygosity). Therefore, the task was to properly identify each of the components independently.

ALGORITHM

The core of VTag is a probability model called Conditional Random Fields (CRFs) (Lafferty *et al*., 2001). These models are convenient because they allow us to combine the effects of many potentially informative features and have previously been successfully used for other biomedical named entity taggers (McDonald and Pereira, 2004). CRFs model the conditional probability of a tag sequence given an observation sequence:

$$P(\mathbf{T}|\mathbf{O}) = \frac{e^{\sum_i \sum_j \lambda_i f_i(t_j, t_{j-1}, \mathbf{O})}}{\sum_{\mathbf{T}'} e^{\sum_i \sum_j \lambda_i f_i(t'_j, t'_{j-1}, \mathbf{O})}},$$

where $\mathbf{O}$ is an observation sequence, in our case a sequence of tokens in the abstract, and $\mathbf{T} = t_1, t_2, \ldots, t_n$ is a corresponding tag sequence in which each tag labels the corresponding token with one of TYPE, LOCATION, INITIAL-STATE, ALTERED-STATE and OTHER. CRFs are log-linear models based on a set of feature functions, $f_i(t_j, t_{j-1}, \mathbf{O})$ that map predicates on observation/tag-transition pairs to binary values. Each feature has an associated weight, λ_i, that measures its effect on the overall choice of tags. These models are convenient because they allow us to combine the effects of many potentially informative features. For example, we may want to include the feature:

$$f_i(t_j, t_{j-1}, \mathbf{O}) = \begin{cases} 1.0 & t_j = \textbf{TYPE}, t_{j-1} = \textbf{TYPE} \\ & \mathbf{o}_j = \textbf{mutation}, \mathbf{o}_{j-1} = \textbf{point} \\ 0 & \text{o.w.} \end{cases}$$

Good features represent informative associations between observation predicates and their corresponding labels, and should receive high weights. For instance, the above feature would most likely receive a high weight, since it is very good evidence that a token is a variation type if the token is the word 'mutation', the previous token was 'point' and the previous token was also part of a variation type. To define the set of features, first we created a set of observation predicates. The set of observation predicates used by the system include word, character-*n*-gram and orthographic predicates such as capitalization. For domain-specific predicates we created a number of regular expressions. For example we included the regular expression:

chr|chromosome [1–9]|1[0–9]|2[0–2]|X|Yp|q

to indicate tokens that might be part of a variation location. If a contiguous set of input tokens match a regular expression (i.e. 'chr 17 p' would match the above expression), then that predicate is set to true for all tokens that participated in the match. All predicates were then applied over all labels and a token window of (−1, 1) to create the final set of features. In total, there were 27 feature types with a total of 63 421 unique features (a complete list is available in our documentation).

The CRF parameters (feature weights) λ_i are trained to maximize the penalized log-likelihood of the training data $\Im$:

$$\sum_{(\mathbf{T},\mathbf{O}) \in \Im} \log P(\mathbf{T}|\mathbf{O}) - \sum_i \frac{\lambda_i^2}{\sigma^2},$$

where the second term controls overfitting by penalizing the large weights that would otherwise arise from rarely observed features. This maximization has no closed form solution, but it can be done efficiently with suitable convex optimization methods (Sha and Pereira, 2003). Given a trained model, the optimal tag sequence for new examples is found with the Viterbi algorithm (Rabiner, 1993). We used the MALLET toolkit (McCallum, 2002, http://mallet.cs.umass.edu) implementation of CRF as the core of our model.

RESULTS

Our training set abstracts were selected from MEDLINE as being relevant to populating a database with facts of the form 'gene X with variation event Y is associated with malignancy Z'. VTag was trained and tested using a corpus of 545 abstracts manually annotated by domain specialists. The abstracts were randomly chosen from a larger corpus identified as containing variation mentions pertaining to cancer. Abstracts were obtained through MEDLINE based upon their PubMedIDs and added to a customized workflow system. Prior to entity tagging, each abstract was tokenized and annotated for part-of-speech. Entity tagging was performed by trained annotators using a locally developed, open-source tool (WordFreak, 2004, http://sourceforge.net/projects/wordfreak). Entity annotators manually identified all mentions of the variation components and labeled each mention with the appropriate tag (type, location, state).

Manual entity annotations for the three variation components were then used to train the variation component tagger. Data, documentation and entity definitions are available by contacting the authors (UPenn Biomedical Information Extraction Group, 2003). Of the 545 abstracts that were annotated, 345 were used as training and development data for the system. The remaining 200 files were used as evaluation data. The tagger took ∼5 h to train on an Intel Xeon 3.2 GHz Linux server. Once trained, VTag can tag a new abstract in under a second.

For evaluation purposes, an entity was considered correctly identified if and only if the predicted and manually labeled tags were identical in both category (e.g. type, location or state) and span (i.e. character positions X through Y). The performance of VTag was calculated according to the following metrics: Precision (number of entities predicted correctly divided by the total number of entities predicted),

Table 1. System performance on evaluation data

Entity	Precision	Recall	*F*-measure
Type	0.8556	0.7990	0.8263
Location	0.8695	0.7722	0.8180
State-Initial	0.8430	0.8286	0.8357
State-Sub	0.8035	0.7809	0.7920
Overall	0.8541	0.7870	0.8192

Recall (number of entities predicted correctly divided by the total number of entities in the text) and *F*-measure [(2∗Precision∗Recall)/(Precision + Recall)]. Performance for individual variation components as well as the overall performance is given in Table 1.

VTag is one of a number of tools under construction as part of an information extraction toolkit called BioSFIER (Biological Software For Information Extraction and Retrieval). Performance should increase as additional documents are annotated and used for training. A long-term objective of our ongoing project is to produce various forms of syntactic and semantic annotation of biomedical text documents (Kulick *et al*., 2003) to aid in information extraction, including the development of algorithms to extract both named entities and events (relationships between entities). VTag serves as the foundation for our development of a variation event tagger useful in recognizing relationships between variation components.

ACKNOWLEDGEMENTS

The authors thank Eric Pancoast for technical assistance; Richard Wooster for corpus provision; and the Penn BIEG for annotations and helpful advice. This work was supported in part by NSF grant ITR 0205448.

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 Lingustics (COLING'2003)*, Saarbrucken, Germany, pp. 201–207.

Kulick,S., Liberman,M., Palmer,M. and Schein,A. (2003) Shallow semantic annotations of biomedical corpora for information extraction. In *Proceedings of the 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 Proceedings of ICML-01*, pp. 282–289.

McCallum,A.K. (2002) MALLET: a machine learning for language toolkit.

McDonald,R. and Pereira,F. (2004) Identifying gene and protein mentions in text using conditional random fields. In *A Critical Assessment of Text Mining Methods in Molecular Biology workshop, 2004*.

Rabiner,L. (1993) A tutorial on hidden Markov models and selected applications in speech recognition. In Waibel,A. and Lee,K.F. (eds), *Readings in Speech Recognition*. Morgan Kaufmann Publishers, San Francisco, CA, pp. 267–296.

Sha,F. and Pereira,F. (2003) Shallow parsing with conditional random fields. In *Proceedings of HLT-NAACL 2003*. Association for Computational Linguistics, pp. 213–220.

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.

Wang,J.Y., Lian,S.T., Chen,Y.F., Yang,Y.C., Chen,L.T., Lee,K.T., Huang,T.J. and Lin,S.R. (2002) Unique K-ras mutational pattern in pancreatic adenocarcinoma from Taiwanese patients. *Cancer Lett.*, **180**, 153–158.

Yu,H., Hatzivassiloglou,V., Rzhetsky,A. and Wilbur,W.J. (2002) Automatically identifying gene/protein terms in MEDLINE abstracts. *J. Biomed. Inform.*, **35**, 322–330.