Identifying and extracting malignancy types in cancer literature

Yang Jin[1], Ryan T. McDonald [2], Kevin Lerman [2], Mark A. Mandel [4], Mark Y. Liberman [2, 4], Fernando Pereira [2], R. Scott Winters [3], Peter S. White [1, 3, ‡]

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

[‡] To whom correspondence should be addressed.

ABSTRACT

Summary: MTag is an application for identifying and extracting clinical descriptions of malignancy presented in text. The application uses the machine learning technique Conditional Random Fields and incorporates domain-specific features. MTag was tested with 1,010 training and 432 evaluation documents pertaining to cancer genomics. Our experiments resulted in 0.85 precision, 0.82 recall, and 0.83 F-measure on the evaluation set.
Availability: The software is available at http://bioie.ldc.upenn.edu/index.jsp
Contact: yajin@mail.med.upenn.edu

INTRODUCTION

The biomedical literature collectively represents the acknowledged historical perception of biological and medical concepts, including findings pertaining to cancer research. However, the rapid proliferation of this information makes it increasingly difficult for researchers and clinicians to peruse, query, and synthesize it for biomedical knowledge gain. Automated information extraction methods, which have recently been increasingly concentrated upon biomedical text, can assist in the acquisition and management of this data. Much of this effort has focused upon molecular object (entity) classes, including gene/protein names and protein interactions, and entity recognition algorithms for these tasks have improved considerably in the last few years (Leek 1997, Collier *et al.* 2000, Tanabe and Wilbur 2002, Yu *et al.* 2003, GENIA 2004, Temkin *et al.* 2003, Huang *et al.* 2004). We recently extended this focus to include genomic variations (McDonald *et al.* 2004). Although there have been efforts to apply automated entity recognition to the identification of phenotypic and disease objects (Friedman et al. 1995; Hahn et al., 2000), these systems often do not perform as well as those utilizing more recently evolved machine-learning techniques for such tasks as gene recognition. However, medical entity class recognition is an important prerequisite for utilizing structured text information to improve clinical applications.

To determine the feasibility of efficiently capturing disease descriptions, we describe here an algorithm for automatically recognizing a specific disease entity class: malignant disease labels. This algorithm, MTag, is based upon a Conditional Random Fields model successfully employed in recognizing other biomedical entities (McDonald and Pereira 2004, McDonald *et al.* 2004). The algorithm considers a large number of syntactic and semantic features of the text surrounding each putative mention. MTag directly takes MEDLINE-formatted abstracts from PubMed as input. The output consists of a text file containing a list of identified malignancy types and an HTML file displaying color-coded malignancy types highlighted in the original abstract text. To the best of our knowledge, MTag is the first direct effort at automated literature extraction of a specific disease class. Immediate applications of this algorithm include automation-assisted generation of exhaustive vocabularies and subsequent utility for complex query expansion.

TASK

Our task was to develop an automated method that would accurately identify and extract strings of text corresponding to a clinician's or researcher's reference to cancer (malignancy type). Our definition of the extent of malignant type was generally the full noun phrase encompassing a mention of a cancer subtype, such that "neuroblastoma", "localized neuroblastoma", and "primary extracranial neuroblastoma" were considered to be distinct malignant type mentions. Attached prepositional phrases, such as "cancer <of the lung>", were not allowed, as these constructions often denoted ambiguity as to exact type. Within these confines, the task included identification of all variable descriptions of particular malignant types, such as the forms "squamous cell carcinoma" (histological observation) or "lung cancer" (anatomical location), both of which are underspecified forms of "lung squamous cell carcinoma".

METHOD

In order to train and test the tagger with both depth and breadth, we combined two corpora, for testing. The first concentrated upon a specific malignancy (neuroblastoma) and consisted of 1000 randomly selected abstracts identified by querying PubMed with the query terms "neuroblastoma" and "gene". Of these, 158 abstracts were manually eliminated if they appeared to be non-topical, had no abstract body, or were not written in English. The second corpus consisted of 600 abstracts previously selected as likely containing gene mutation instances for genes commonly mutated in a wide variety of malignancies, and for which genomic and malignant annotations had been previously performed manually. These sets were combined to create a single corpus of 1442 abstracts. This set was manually annotated for tokenization, part-of-speech assignments (Kulick *et al*. 2003, Upenn Biomedical Information Extraction Group, 2004), and malignant type named entity recognition, the latter in strict adherence to our pre-established entity class definition (http://www.cis.upenn.edu/~mamandel/annotators/ent-genrules.html). Dual pass annotations were performed on all documents by experienced annotators with biomedical knowledge, and discrepancies were resolved through forum discussions. A total of 7303 malignant type mentions were identified in the document set.

Based on the manually annotated data, an automatic malignancy type tagger (MTag) was developed using the probability model Conditional Random Fields (CRFs) (Lafferty *et al*., 2001). We have previously demonstrated that this model yields state-of-the-art accuracy for recognition of biomedical named entity classes (McDonald and Pereira 2004, McDonald *et al*. 2004). CRFs model the conditional probability of a tag sequence given an observation sequence. We denote that $\boldsymbol{O}$ is an observation sequence, or a sequence of tokens in the text, and $\boldsymbol{t}$ is a corresponding tag sequence in which each tag labels the corresponding token with either *Malignancy Type* (meaning that the token is part of a malignancy type mention) or *Other*. CRFs are log-linear models based on a set of feature functions, $f_i(t_j, t_{j-1}, \boldsymbol{O})$, which map predicates on observation/tag-transition pairs to binary values. As shown in the formula below, the function value is 1.0 when the tag sequence is malignancy type; otherwise (o.w.) it is 0. A particular advantage of this model is that it allows the effects of many potentially informative features to be simultaneously weighed. Consider, for example, the following feature:

$$f_i(t_j, t_{j-1}, O) = \begin{cases} 1.0 & t_j = \text{Malignancy Type}, t_{j-1} = \text{Malignancy Type} \\ & O_j = \text{cancer}, O_{j-1} = \text{lung} \\ 0 & O.W. \end{cases}$$

This feature represents the probability of whether the token "cancer" is tagged with label malignant type given the presence of "lung" as the previous token. Features such as this would likely receive a high weight, as they represent informative associations between observation predicates and their corresponding labels. A set of observation predicates, including word and

character-*n*-gram characterizations and orthographic predicates (e.g. capitalization patterns) were defined. In addition, we created biomedically-derived predicates, including regular expression patterns (e.g. the suffix –oma) and specified lexicons [e.g. terms from the National Cancer Institute (NCI) neoplasm ontology.] All predicates were then applied over all labels, applying a token window of (-1, 1) to create the final set of features. In total there were six feature types together with 80,294 unique features. The MALLET toolkit (McCallum 2002) was used as the implementation of CRFs to build our model.

RESULTS

Manually annotated texts from the corpus of 1442 MEDLINE abstracts were used to train and evaluate MTag. MTag was tested with a randomly selected 1,010 (70%) training and 432 (30%) evaluation documents pertaining to cancer genomics. The tagger took approximately 6 hours to train on a 733 MHz PowerPC G4 with 1 GB SDRAM Mac server. Once trained, MTag can tag a new abstract in a matter of seconds.

For evaluation purposes, manual annotations were treated as gold-standard files (100% annotation accuracy). The evaluation set of 432 abstracts comprised 2,031 sentences containing malignant type mentions and 3,752 sentences without mentions, as determined by manual assessment of entity content. The predicted malignancy type mention was considered correctly identified if, and only if, the predicted and manually labeled tags were exactly the same in content and both boundary determinations. The performance of MTag was calculated according to the following metrics: Precision (number of entities predicted correctly divided by the total number of entities predicted), Recall (number of entities predicted correctly divided by the total number of entities identified manually), and F-measure ((2*Precision*Recall)/(Precision+Recall)). Our experiments resulted in 0.85 precision, 0.82 recall, and 0.83 F-measure on the evaluation set. Additionally, the two subset corpora (neuroblastoma-specific and gene-specific) were tested separately. The tagger performed with higher accuracy with the more narrowly defined (neuroblastoma) corpus than with the corpus more representative for various malignancies (gene-specific). The neuroblastoma corpus performed with 0.88 precision, 0.87 recall, and 0.88 F-measure, while the gene-specific corpus performed with 0.77 precision, 0.69 recall, and 0.73 F-measure. These results likely reflect the increased challenge of identifying malignant type mentions in a document set demonstrating a more diverse collection of mentions.

Performance of the tagger relative to a baseline system that could be easily employed by a typical research group was also evaluated. For the baseline system the NCI neoplasm ontology, a term list of 5,555 malignant types, was used as a lexicon to identify malignancy type mentions. Lexicon terms were individually queried against text by exact string matching. A subset of 39 abstracts randomly selected from the testing set, which together contained 202 malignancy type mentions, were used to compare the automated tagging and baseline results. The tagger identified 190 of the 202 mentions correctly (94.1%), while the NCI list identified only 85 (42.1%), all of which were also identified by the tagger. Analysis of the results suggested that the major deficiencies of the lexical approach were the inability to identify minor variations in spelling and form (e.g. neuroblastomas), and the inability to identify acronyms (e.g. AML).

MTag has been engineered to directly accept downloaded files from PubMed and formatted in MEDLINE format as input, and to output text and HTML file versions of the tagger results. The text file is similar to the input file, except for the identified malignancy types appended at the end. The HTML file shows the original abstract with color-highlighted malignancy types as demonstrated in the following tagged MEDLINE abstract by Bruder *et al.*:

Normal text
Malignancies

PMID: 15316311
Morphologic and molecular characterization of *renal cell carcinoma* in children and young

adults.
A new WHO classification of ***renal cell carcinoma*** has been introduced in 2004. This classification includes the recently described ***renal cell carcinomas*** with the ASPL-TFE3 gene fusion and ***carcinomas*** with a PRCC -TFE3 gene fusion. Collectively, these tumors have been termed Xp11.2 or TFE3 ***translocation carcinomas***, which primarily occur in children and young adults. To further study the characteristics of ***renal cell carcinoma*** in young patients and to determine their genetic background, 41 ***renal cell carcinomas*** of patients younger than 22 years were morphologically and genetically characterized. Loss of heterozygosity analysis of the von Hippel - Lindau gene region and screening for VHL gene mutations by direct sequencing were performed in 20 tumors. TFE3 protein overexpression, which correlates with the presence of a TFE3 gene fusion, was assessed by immunohistochemistry. Applying the new WHO classification for ***renal cell carcinoma***, there were 6 clear cell (15 %), 9 papillary (22 %), 2 chromophobe, and 2 collecting duct ***carcinomas***. Eight ***carcinomas*** showed translocation carcinoma morphology (20 %). One ***carcinoma*** occurred 4 years after a ***neuroblastoma***. Thirteen tumors could not be assigned to types specified by the new WHO classification: 10 were grouped as unclassified (24 %), including a unique ***renal cell carcinoma*** with prominently vacuolated cytoplasm and WT1 expression. Three ***carcinomas*** occurred in combination with ***nephroblastoma***. Molecular analysis revealed deletions at 3p25-26 in one ***translocation carcinoma***, one ***chromophobe renal cell carcinoma***, and one ***papillary renal cell carcinoma***. There were no VHL mutations. Nuclear TFE3 overexpression was detected in 6 ***renal cell carcinomas***, all of which showed areas with voluminous cytoplasm and foci of papillary architecture, consistent with a ***translocation carcinoma*** phenotype. The large proportion of TFE3 " translocation " ***carcinomas*** and "unclassified " ***carcinomas*** in the first two decades of life demonstrates that ***renal cell carcinomas*** in young patients contain genetically and phenotypically distinct tumors with further potential for novel ***renal cell carcinoma*** subtypes. The far lower frequency of ***clear cell carcinomas*** and VHL alterations compared with adults suggests that ***renal cell carcinomas*** in young patients have a unique genetic background.

MTag can be utilized and further explored in various ways. First, when combined with expert evaluation of output, it can help build a vocabulary for all the synonyms of cancer names, which is of great benefit for data integration procedures requiring normalization of malignant types. However, unlike molecular entity classes such as genes, such supervised lists are often not readily available, due in part to the variability in which phenotypic and disease descriptions can be described, and in part to the lack of nomenclature standards in many cases. Secondly, to the best of our knowledge, MTag is the first significant effort to automatically extract entity mentions in a disease-oriented domain. Therefore, this is an important contribution towards a process of identifying and extracting associations between molecular and clinical objects in an automation-centric manner. MTag and its underlying algorithm have been designed to be rapidly adaptable to other biomedical entity classes. Thus, as MTag performs well for extracting malignancy types, this procedure can subsequently be expanded to extract additional disease-oriented information, including clinically-derived observations. Future work will include determining how well similar taggers perform for identifying mentions of malignant attributes with greater (e.g. tumor histology) and lesser (e.g. tumor clinical stage) semantic and syntactic heterogeneity.

ACKNOWLEDGEMENTS

The authors thank members of the University of Pennsylvania Biomedical Information Extraction Group; Kevin Murphy for annotations, discussions and technical assistance; and Richard Wooster for corpus provision. This work was supported in part by NSF grant ITR 0205448.

REFERENCE

Bruder, E., Passera, O., Harms, D., Leuschner, I., Ladanyi, M., Argani, P., Eble, J.N., Struckmann, K., Schraml, P., Moch, H. (2004) Morphologic and molecular characterization of renal cell carcinoma in children and young adults. *American Journal of Surgical Pathology*, 28:1117-1132.

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)*, Saarbrucken, Germany.

Friedman, C., Hripcsak, G., DuMouchel, W., Johnson, S.B., and Clayton, P.D. (1995) Natural language processing in an operational clinical information system. *Natural Language Engineering*, 1:1-28.

GENIA. (2004) http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/

Hahn, U., Romacker, M., Schulz, S. (2000) medSynDiKATe: A natural language system for the extraction of medical information from findings reports. *International Journal of Medical Informatics* 67:63-74.

Huang, M.H., Zhu, X., Hao, Y., Payan, D.G., Qu, K., and Li, M. (2004) Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics,* 20:3604-3612.

Kulick, S., Liberman, M., Palmer, M. and Schein, A. (2003) Shallow semantic annotations of biomedical corora 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.

Leek, T.R. (1997) Information Extraction Using Hidden Markov Models. Dissertation, University of California, San Diego.

McCallum, A.K. (2002) http://mallet.cs.umass.edu/

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.

McDonald, R.T., Winters, R.S., Mandel, M., Jin, Y., White, P.S. and Pereira, F. (2004). An entity tagger for recognizing acquired genomic variations in cancer literature. *Bioinformatics,* 22:3249-3251.

Tanabe, L. and Wilbur, W. (2002) Tagging gene and protein names in biomedical text. *Bioinformatics*, 18:1124-1132.

Temkin, J.M. and Glder, M.R. (2003) Extraction of protein interaction information from unstructured text using a content-free grammar. *Bioinformatics*, 19:2046-2053.

Upenn Biomedical Information Extraction Group (2004) http://bioie.ldc.upenn.edu/

Yu, H., Hatzivassiloglou, V., Rzhetsky, A. and Wilbur, W.J. (2002) Automatically identifying gene/protein terms in MEDLINE abstracts. *Journal of Biomedical Informatics,* 35:322-330.