Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Biomedical Informatics 37 (2004) 393–395
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Guest Editorial

Introduction: named entity recognition in biomedicine

This special issue responds to the increasing interest of the biomedical community in text mining techniques. This is an exciting time for the text processing community, as there is an urgent need for text mining tools and methods in the biomedical domain. The amount of biological literature published daily is growing exponentially. Medline alone contains 14 million abstracts and is a critical source of information for biologists and curators. As these scientists find it essential to search for information in an overabundance of documents, their need for text mining techniques tailored to the biological domain has become apparent.

The focus of this special issue is on named entity recognition (NER) in biomedicine, a fundamental challenge for text mining due to the special problems caused by the complex nature of biological entity recognition, classification, and unique identification. This is a key factor for access to the information stored in literature, as it is the biological entities and their relationships that convey knowledge across scientific articles.

Textual terms (names of genes, proteins, gene products, organisms, drugs, chemical compounds, etc.) are the primary means of scientific communication because they are used in language to represent the concepts in the domain; it would be impossible to "understand" an article or to extract information from it without the precise identification and association of the terms. Biomedical terminology presents a special challenge. It is constantly changing; new terms are rapidly being introduced for each of the organisms being studied, while old ones are discarded (e.g., withdrawn or made obsolete). Biological names are very complex, as they are created and referenced by many different communities. They include an enormous amount of synonyms and variant forms, such as acronyms, and morphological, derivational, and orthographic variants, all of which are used interchangeably in the literature. In addition, many biological terms and their variants are ambiguous. They share their lexical representations with common English words (gene names/symbols, such as *an*, *by*, *can*, and *for*), or with other biomedical terms (gene names, such as *demented*, *white eye*, and *hair loss*). Existing text processing resources typically lack information that can support disambiguation of terms. Also, terminological resources do not address ambiguities related to finer biological classification, such as species information (homologous genes have the same name, but belong to different species). In many cases different disambiguation approaches are needed to link a recognized term to a correct concept. Also, a narrow context may not always be enough to disambiguate a term (e.g., when a protein name is shared between species), and a wider context (e.g., a whole article) may need to be analyzed before terms can be mapped.

The collection of papers in this issue reports on diverse approaches that use a variety of natural language processing, corpus-based and machine learning techniques to recognize, classify, and/or identify biological entities. Recognition involves identifying the boundaries of the name in the text, whereas classification assigns a semantic class to the entity based on an appropriate ontology, and identification maps the term to a normalized form or to a unique identifier.

The paper by Morgan et al. [1] reports on a series of experiments related to the application of natural language processing as a tool to aid in curation of the FlyBase database. They used Flybase resources, and a combination of techniques, such as pattern matching, tokenization, HMM-based tagging, disambiguation heuristics, etc., to automatically generate large quantities of high-quality training data to support the automatic learning of a gene name recognizer. The generation of normalized gene lists is also explored using simple pattern matching and an HMM gene name tagger.

Zhang and colleagues [2] adapt an HMM named entity recognizer to the biomedical domain via a rich feature set consisting of orthographic, morphological, part-of-speech, and semantic trigger features. These features are integrated via a HMM with back-off modeling. In addition, they propose methods for recognizing

1532-0464/$ - see front matter
doi:10.1016/j.jbi.2004.08.011

biomedical abbreviations and cascading (i.e., nested) named entities. Their treatment of the cascading phenomenon is novel as they recognize both the nested and the longest named entities. In their work, they propose two approaches for recognizing cascaded names: a post-processing rule-based approach and an HMM-based approach.

Variations of character-level orthographic features and part-of-speech (POS) features on the performance of NERs are examined by Collier and Takeuchi [3]. Their experiments, which are based on support vector machines (SVMs), revealed that orthographic features outperformed POS features. The reasons that POS features appear to be less useful than orthography are due to the complex relationship between name boundaries, local syntactic ambiguities, and class semantics. In addition, they demonstrate that the combination of orthographic features and POS degrades the overall performance of NERs slightly.

Lee et al. [4] present a two-phase named entity recognizer based on SVMs, which consists of two subtasks: a boundary identifier and a semantic classifier of named entities. This separation of the NER task allows the use of the appropriate SVM classifier and relevant features for each subtask, resulting in a reduction of computational complexity and improvement in performance. A hierarchical classification method is employed for semantic classification that utilizes 22 semantic classes that are based on the GENIA ontology [5].

An automatic method for mining collocates (i.e., two or more words that occur together much more frequently than by chance) in the literature is proposed by Hou and Chen [6]. They focus on collocations associated with gene and protein names, and use the extracted collocates to improve the precision rate of protein and gene name recognition. In addition, they integrate the results of multiple NERs, such as Yapex [7], KeX [8], ABGene [9], and Idgene [10], to improve the recall rates. The combination of filtering and integration strategies increased the performance of the NER.

Novel techniques for boosting the performance of dictionary-based protein name recognition are suggested by Tsuruoka and Tsujii [11]. They propose two alternative methods to tackle the problem of low recall due to spelling variations. One method uses approximate string matching, where similarity between two strings is computed based on an edit distance. What is interesting about their method is that the cost for individual operations varies depending on the letter being operated on (e.g., substitution of an alphabetic character costs more than substitution of a dash or a number). An alternate method, which is more efficient, involves expanding the dictionary in advance using a probabilistic variant generator. A method to filter out false positives is also presented, which is based on use of a naïve Bayes classifier.

The use of morphological analysis in protein name recognition to overcome problems such as boundary disagreement is proposed by Yamamoto and colleagues [12]. To overcome boundary disagreement that is caused by tokenization ambiguity, they apply techniques borrowed from Japanese (nonsegmented language) morphological analysis. The authors show that their augmented preprocessing improves the performance of protein name recognition over conventional preprocessing.

Spasic and Ananiadou [13] examine term classification for the task of ontology management, where it is of interest to automatically augment an ontology with novel terms. A genetic algorithm is used in order to refine verb selectional preferences and to assign classes to domain-specific verbs. The class of a newly discovered term is suggested depending on its co-occurrence with a domain-specific verb, as well as a similarity measure to known terms with established term-class relationships.

The topic of term classification (independently from the task of term identification) is examined by Torii and colleagues [14]. They focus on different sources of information that can be used for classification and report on their effectiveness. They apply machine learning methods to build a classifier, and they use both name-internal features (e.g., suffixes) and name-external features (e.g., contextual information) for the classification task.

To conclude this issue, state-of-the art approaches to term identification are reviewed by Krauthammer and Nenadic [15]. The paper features an extensive list of work published in the domain. It analyzes the process of identifying terms through three steps: term recognition, term classification, and term mapping, which in some cases can be merged. For each step, the main approaches and general trends, along with the major problems, are discussed. Also, by identifying various challenges that term identification is still faced with, the review tries to delineate directions for future work in the field.

References

[1] Morgan AA, Hirschman L, Colosimo M, Yeh A, Colombe J. Gene name identification and normalization using a model organism database. J Biomed Inform 2004;37(6):396–410.
[2] Zhang J, Shen D, Zhou G, Su J, Tan CL. Enhancing HMM-based biomedical named entity recognition by studying special phenomena. J Biomed Inform 2004;37(6):411–22.
[3] Collier N, Takeuchi K. Comparison of character-level and part of speech features for name recognition in bio-medical texts. J Biomed Inform 2004;37(6):423–35.
[4] Lee KJ, Hwang YS, Kim S, Rim HC. Biomedical named entity recognition using two phase model based on SVMs. J Biomed Inform 2004;37(6):436–47.
[5] Ohta, T, Tateisi Y, Kim J, Mima H, Tsujii J-I. The GENIA Corpus: an annotated research abstract corpus in molecular biology domain. In: The 2nd International Conference on Human Language Technology. p. 82–6.

[6] Hou WJ, Chen HH. Enhancing performance of protein and gene name recognizers with filtering and integration strategies. J Biomed Inform 2004;37(6):448–60.
[7] Olsson F, Eriksson G, Franzen K, Asker L, Liden P. Notions of correctness when evaluating protein name taggers. In: Proceeedings of the 19th International Conference on Computational Linguistics 2002. p. 765–71.
[8] Fukuda K, Tsunoda T, Tamura A, Takagi T. Toward information extraction: identifying protein names from biological papers. Pac Symp Biocomput 2003:403–14.
[9] Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. Bioinformatics 2002;18(8):1124–32.
[10] Fan JW. Information retrieval and extraction for the Chinese Gene Variation Database (CGVdb). Unpublished Master's Thesis; 2003.
[11] Tsuruoka Y, Tsujii J. Improving the performance of dictionary-based approaches in protein name recognition. J Biomed Inform 2004;37(6):461–70.
[12] Yamamoto K, Kudo T, Konagaya A, Matsumoto Y. Use of morphological analysis in protein name recognition. J Biomed Inform 2004;37(6):471–82.
[13] Spasic I, Ananiadou S. Using automatically learnt verb selectional preferences for classification of biomedical terms. J Biomed Inform 2004;37(6):483–97.
[14] Torii M, Kamboj S, Vijay-Shanker K. Using name-internal and contextual features to classify biological terms. J Biomed Inform 2004;37(6):498–511.
[15] Krauthammer M, Nenadic G. Term identification in the biomedical literature. J Biomed Inform 2004;37(6): 512–26.

Sophia Ananiadou
School of Computing, Science and Engineering
University of Salford
Machester M5 4WT, UK

Carol Friedman*
Department of Biomedical Informatics
Columbia University
622 West 168 St., VC Bldg., 5th Floor
New York, NY 10032, USA
E-mail address: friedman@dbmi.columbia.edu

Jun'ichi Tsujii
Department of Computer Science
Graduate School of Information Science and Technology
University of Tokyo
Hongo Bunkyu, Tokyo 113-0033, Japan

Available online 8 October 2004

* Corresponding author. Fax: +1 212 305 3302.