Biodiversity Informatics, 2, 2005, pp. 79-82

TAXONGRAB: EXTRACTING TAXONOMIC NAMES FROM TEXT

DREW KONING[1], INDRA NEIL SARKAR[1,2], AND THOMAS MORITZ[1,3]
[1]Divisions of Library Services and [2]Invertebrate Zoology,
American Museum of Natural History, New York, NY 10024 USA
[3]E-mail: tmoritz@amnh.org

Abstract.—Identification of organism names in biological texts is essential for the management of archival resources to facilitate comparative biological investigation. Because organism nomenclature conforms closely to prescribed rules, automated techniques may be useful for identifying organism names from existing documents, and may also support the completion of comprehensive indices of taxonomic names; such comprehensive lists are not yet available. Using a combination of contextual rules and a language lexicon, we have developed a set of simple computational techniques for extracting taxonomic names from biological text. Our proposed method consistently performs at greater than 96% Precision and 94% Recall, and at a much higher speed than manual extraction techniques. An implementation of the described method is available as a Web based tool written in PHP. Additionally, the PHP source code is available from SourceForge: http://sourceforge.net/projects/taxongrab, and the project website is http://research.amnh.org/informatics/taxlit/apps/.

Key words.—Named Entity Recognition; Taxonomic Name Extraction

INTRODUCTION

New and revised biological names are often embedded within the conventional biological and medical literature. While some taxonomic names are available via popular indexing services (e.g., *Zoological Record*), complete names data are only available in the form of printed articles. As the full legacy of this printed literature is digitized, automated techniques will be in demand to assist with the extraction and indexing of these, named entities.

A range of computational techniques, categorized as Named Entity Recognition (NER) techniques, exist for identifying named entities (Cunningham et al. 2002). NER seeks to locate and classify atomic elements in text into predefined categories. For example, NER techniques have shown utility in identifying gene names from biomedical articles (Krauthammer et al. 2004). Organism names represent another common named entity that may be usefully extracted, particularly in cases where the literature consists of information pertaining to organism biology or biodiversity (Soberón et al. 2004).

Organism names within taxonomy generally occur in natural language text as sequences of two or three words, called "binomen" and "trinomen" taxonomic names, respectively. Taxonomic names also follow a prescribed set of linguistic and contextual rules (Linnaeus 1753): The scientific name of an organism is written in either Latin or Greek. Genus name precedes species name, and is capitalized. Species name, and any subsequent subspecies, variant, or strain names are written in lower case. As organisms are discovered and described in scientific literature, these rules are mandated and prescribed by International Commissions (e.g., ICZN and ICBN for Zoological or Botanical organisms, respectively).

A commonly used approach in NER is to create dictionaries of terms that can later be referenced to identify known named entities (Petasis et al. 2000). This approach has limited success, as it requires the term to pre-exist in the dictionary of terms. In order to create a comprehensive dictionary of taxonomic names for currently recognized species, one would currently need to craft a dictionary consisting of 1.5-1.8 million items (Wilson 2003). With the number of new organisms identified increasing, in pace with advances in collection and description methods, this number could, ultimately, range from 30 to 100 million (Wilson 2003). As a result, a comprehensive and current catalogue of taxonomic names will be needed to create reliable look-up and indexing algorithms.

The linguistic and contextual nature of taxonomic names, as dictated by Linnaean rules, enables the development of computational tools that can extract names from natural language text. Since taxonomic names are most typically – though not exclusively – derived from Greek or Latin, filters can be used to separate taxonomic name candidates using a language-specific lexicon (e.g., a lexicon of English words). Finally, after taxonomic name candidates have been identified, the Linnaean conventions for capitalization (i.e., Genus name is capitalized, followed by lowercase species and subspecies names) can be used to identify taxonomic name entities.

Here, we explore the applicability of NER methods for identifying taxonomic names from digitized texts. Using a lexicon of words that we have compiled from existing English lexicons, we assess the efficacy of our method for extracting taxonomic names. We evaluate the proposed technique with respect to manually identified taxonomic names using a Web-based interface. We conclude with some discussion of how this approach, which we term "TaxonGrab," can be used in the design and development of future taxonomic literature organization systems.

BUILDING TAXONGRAB

The premise of the TaxonGrab approach is that taxonomic names can be identified from natural language text using a combination of taxonomic nomenclature rules and a lexicon of non-taxonomic terms. We composed a lexicon of English words by combining the terms from the WordNet® (Fellbaum 1998) and SPECIALIST (McCray et al. 1993) lexicons. The WordNet lexicon contains common words that are associated with many facets of the English language. The SPECIALIST lexicon, which is part of the United States National Library of Medicine's Unified Medical Language System® (Lindberg et al. 1993), consists of both common English and biomedical vocabulary words, including spelling variants and inflected forms (e.g., plural).

An important consideration in building the language lexicon from existing resources was that many parts of taxonomic names have become part of the common language, and are therefore included in a complete lexicon; for example, *coli* (e.g., the second lexical unit of the binomen *Escherichia coli*) is included in comprehensive language lexicons. Therefore, in order to create a lexicon that did not contain any taxonomic terms (i.e., words used to describe genus, species, or subspecies), we manually culled words from a list of taxonomic terms drawn from popular taxonomic resources. Specifically, we removed terms from our lexicon that were associated with 362,430 taxonomic names in either the National Center for Biotechnology Information (NCBI) Taxonomy, the Integrated Taxonomic Information System (ITIS), or the German Collection of Microorganisms and Cell Cultures (Deutsche Sammlung von Mikroorganismen und Zellkulturen [DSMZ]). The resulting lexicon consisted of 258,783 words.

A script was written in PHP that used the resulting lexicon. The script isolates sets of two or three consecutive words that are not in the lexicon, as candidate names. These taxonomic name candidates are then validated according to the capitalization rules of Linnaean nomenclature. Additionally, strings involving a capital letter followed by a period and then at least one word that is not in the lexicon – a conventional form of abbreviation for binomens or trinomens that have been previously cited in a work (e.g., *Escherichia coli* is often abbreviated as *E. coli*) – are also reported as taxonomic name candidates. The script also looks for more complex taxonomic formatting rules – such as *variants* preempted with *var.*, *subspecies* preempted with *subsp.*, or parentheses that are often used to indicate sub-genus or author names. All of these rules were implemented in the script using regular expressions.

A Web interface was designed whereby users can enter text or upload text files. The interface then returns the list of taxonomic name candidates that are found using the TaxonGrab method. The interface allows one to enter text either by entering it directly, through uploading a text file, or specifying a Web location.

EVALUATION

The Web interface was used to examine a number of documents, consisting of archived publications, which were digitized using standard optical scanning and optical character recognition techniques (OCR. Using an off-the-shelf software package; Abbyy FineReader© that purports 97% accuracy.

As a test corpus, we used the Volume 1 of "*The Birds of the Belgian Congo"* by *James Paul Chapin (published in four parts in the series: Bulletin of the American Museum of Natural History, 1932-1954).* This corpus consists of 5000 pages that contain over 8000 taxonomic names. TaxonGrab was used via the Web interface to extract names from the corpus. The extracted taxonomic names were then compared to a list of taxonomic names that had been manually identified by a team of experts. The results were then assessed using Precision and Recall values. Precision, or the correctness of the reported taxonomic names, is defined as ratio of correct taxonomic names (TP) to the sum of correct and false taxonomic names (TP+FP): TP/(TP+FP). Recall, or the ability to retrieve taxonomic names, is defined as the ratio of the sum of correct taxonomic names (TP) to the sum of correct and missed taxonomic names (TP+FN): TP/(TP+FN). Compared to the manually extracted taxonomic names, TaxonGrab consistently identified taxonomic names with greater than 96% Precision and 94% Recall from the documents examined. Errors arose mainly from OCR errors, manuscript typos, and the few common English words that are also used as scientific names, which had not been addressed in the lexicon creation. With respect to the speed of extraction, the manual extraction was reported to have taken 80 hours, while the automated method took approximately 330 seconds.

DISCUSSION AND FUTURE DIRECTIONS

With the many advances in biological collection and description techniques, comprehensive and current catalogues of taxonomic names will increasingly be needed to support automated lexical lookup systems that are designed for organizing and aggregating textual data. Currently, with only a small fraction of known organisms named, our taxonomic name catalogues are incomplete. Testing just the three resources for taxonomic names described in this study, we found that there was only partial overlap between existing resources. This is probably attributable to differences in foci of the different catalogues – for example, while NCBI taxonomy is mostly concerned with organisms that are described in MEDLINE and have some biomedical significance, ITIS is more concerned with describing organisms in the context of governmental regulation and biodiversity information. At present, there seems insufficient investment in reconciliation of compiled names between biodiversity and biomedical resources. However, there are some links between resources, for

example, for ITIS taxonomic names there are links to appropriate NCBI taxonomy entries. However, there is not yet a centralized list containing all taxonomic names as they are identified and described. Because taxonomic names generally (with the exception of many virus names) follow a set of prescribed syntax and linguistic rules, it is possible to create automated techniques to extract taxonomic names from literature resources. Here, we have developed the TaxonGrab method, which leverages the linguistic and syntactic properties of taxonomic names.

Taxonomic names can be useful as index terms when organizing large sets of literature. To that end, the TaxonGrab method can be used to extract all the taxonomic names associated with documents, both legacy and prospective, which are subsequently used to organize them. Because TaxonGrab does not rely on a particular taxonomic name catalogue, it seems an efficient tool in extracting and compiling new organism names for inclusion in suitable resources. In this way, TaxonGrab may be a tool that can be used for curating and updating taxonomic name catalogues.

It is also possible that TaxonGrab can assist the editing of digitally captured taxonomic literature. Because of their idiosyncratic properties, taxonomic names may not be easily treatable by normal corrective measures (standard dictionaries and spell-checkers) in OCR or other capture technologies. TaxonGrab may prove to be instrumental in rapidly identifying names for automated or semi-automated methods of proof reading and editing.

We implemented TaxonGrab as a Web-based interface written in PHP. The Web interface enables one to search for taxonomic names that may exist in one of three forms (text, text file, Web site). The source code and associated files are available for free download and can be modified for particular needs. Using the TaxonGrab principle, larger queue application (for example, if implemented as a Perl script), one could search through and organize a large set of documents for taxonomic names. This could be a useful utility that can address a number of important questions, such as "*What taxa are described or mentioned in a particular corpus?*" Subsequent tools could be designed that organize the taxonomic names identified into an ontology, whereby one could track and register organism name changes. While there are implications of a taxonomic name ontology within existing resources (e.g., NCBI Taxonomy is organized into a hierarchy of terms, that are also linked into larger upper-level ontologies), to date there are no specific projects focused on organizing taxonomic names and name changes according to an ontological framework. To that end, we are exploring the automated creation of hierarchic taxonomic name lists. As a list of taxonomic names is updated, we will be able to track the types and numbers of various organisms that are described in various different types of corpora. For example, we could consider the difference between what types of organisms are discussed in biodiversity resources versus biomedical resources. Identifying organisms that are described in both can be used as a means to unify knowledge in some areas (e.g., Medical Entomology). In contrast, identifying organisms that are different between biodiversity resources and biomedical resources will highlight key differences between research foci, but may underscore the importance of studying other related organisms of the same genus for clues (e.g., *Drosophila* studies may want to consider species besides just *melanogaster*).

TaxonGrab was designed to extract taxonomic names from English texts, as reflected by the large English lexicon that was constructed. However, our Web interface currently enables one to search for taxonomic names in, Spanish, German, and French documents using very basic dictionaries (compiled from the WinEdit text editor language dictionaries). We anticipate development of non-English versions of TaxonGrab using available non-English lexicons resources (e.g., EuroWordNet (Vossen 1998)).

CONCLUSION

The identification of taxonomic names within published literature can help guide comparative biological investigations. Here, we have proposed a Named Entity Recognition technique, TaxonGrab, which is based upon the systematic nomenclature rules conventionally used for taxonomic nomenclature in scientific publications. We believe that this method may show general utility for indexing documents containing embedded taxonomic names.

ACKNOWLEDGMENTS

The authors thank members of the project team (National Science Foundation: IIS-0241229) for their comments towards the work described. The authors especially thank Norm Johnson, Donat Agosti, and Liz Nichols for their assistance with the manual extraction of taxonomic names. INS is partially supported by National Science Foundation: DBI-0421604 and the Lewis B. & Dorothy Cullman Program for Molecular Systematics. We also thank researchers from the Marine Biological Laboratory at Woods Hole, Massachusetts for their comments regarding the development of other extraction tools based on TaxonGrab.

LITERATURE CITED

Chapin, J. 1932. The birds of the Belgian Congo: Part 1. Bulletin of the American Museum of Natural History: 65. American Museum of Natural History, New York.

Chapin, J. 1939. The birds of the Belgian Congo: Part 2. Bulletin of the American Museum of Natural History: 75. American Museum of Natural History, New York.

Chapin, J. 1953. The birds of the Belgian Congo: Part 3. Bulletin of the American Museum of Natural History: 75A. American Museum of Natural History, New York.

Chapin, J. 1954. The birds of the Belgian Congo: Part 4. Bulletin of the American Museum of Natural History: 75B. American Museum of Natural History, New York.

Cunningham, H., D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP Tools and Applications. Association for Computational Linguists, Philadelphia.

Fellbaum, C. 1998. WordNet: An Electronic Lexical Database. Cambridge, MIT Press.

Krauthammer, M. and G. Nenadic. 2004. Term identification in the biomedical literature. Journal of Biomedical Informatics 37:512-26.

Lindberg, D. A., B. L. Humphreys, and A. T. McCray. 1993. The Unified Medical Language System. Methods of Information in Medicine 32:281-91.

Linnaeus, C. 1753. Species Plantarum, Stockholm. Sweden.

McCray, A. T., A. R. Aronson, A. C. Browne, T. C. Rindflesch, A. Razi, and S. Srinivasan. 1993. UMLS knowledge for biomedical language processing. Bulletin Medical Library Association 81:184-94.

Morgan, A. A., L. Hirschman, M. Colosimo, A. S. Yeh, and J. B. Colombe. 2004. Gene name identification and normalization using a model organism database. Journal of Biomedical Informatics 37:396-410.

Petasis, G., A. Cucchiarelli, P. Velardi, G. Paliouras, V. Karkaletsis, and C. D. Spyropoulos. 2000. Automatic adaptation of Proper Noun Dictionaries through cooperation of machine learning and probabilistic methods. Proceedings of the Association for Computing Machinery: Special Interest Group on Information Retrieval, Athens, Greece.

Soberon, J. and A. T. Peterson. 2004. Biodiversity informatics: managing and applying primary biodiversity data Philosophical Transactions of the Royal Society of London B, 359:689-698

Vossen, P. 1998. Euro WordNet: A Multilingual Database with Lexical Semantic Networks. Norwell, Kluwer Academic Publishers.

Wilson, E. O. 2003. The encyclopedia of life. Trends in Ecology and Evolution 18:77-80.