Journal of Informetrics 9 (2015) 455–465

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Identifying entities from scientific publications: A comparison of vocabulary- and model-based methods

Erjia Yan*, Yongjun Zhu

College of Computing and Informatics, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA

ARTICLE INFO

Article history:
Received 5 November 2014
Received in revised form 22 April 2015
Accepted 22 April 2015

Keywords:
Entity extraction
Vocabulary
Dictionary
Conditional random fields
Content-aware

ABSTRACT

The objective of this study is to evaluate the performance of five entity extraction methods for the task of identifying entities from scientific publications, including two vocabulary-based methods (a keyword-based and a Wikipedia-based) and three model-based methods (conditional random fields (CRF), CRF with keyword-based dictionary, and CRF with Wikipedia-based dictionary). These methods are applied to an annotated test set of publications in computer science. Precision, recall, accuracy, area under the ROC curve, and area under the precision-recall curve are employed as the evaluative indicators. Results show that the model-based methods outperform the vocabulary-based ones, among which CRF with keyword-based dictionary has the best performance. Between the two vocabulary-based methods, the keyword-based one has a higher recall and the Wikipedia-based one has a higher precision. The findings of this study help inform the understanding of informetric research at a more granular level.

1. Introduction

Publication data embody the very essence of humans' scientific and technological advances. These data have been continuously examined through multidisciplinary efforts. Citation-based methods have traditionally been employed to assess research impact; modern statistical methods have employed various bibliometric networks to cluster research specialties, detect author communities, and identify research topics. While these efforts have revealed patterns of scholarly communication, elucidated the scientific workforce, determined mechanisms of impact assessment, and addressed a slew of issues related to disciplinarity and interdisciplinarity, they were largely driven by the analysis of existing publication metadata (e.g., authors, titles, journals, and references). Consequently, we have limited understanding of the ways to analyze the content of individual papers. Moreover, because knowledge is more effectively expressed through unstructured or semi-structured contents, such as titles, abstracts, keywords, or even full-text, we have yet to find out how to use the contents to examine knowledge production and innovation. Therefore, the current study intends to tackle the complexity of unstructured and semi-structured contents, with a focus on detecting entities – expressions in the contents that convey research-relevant information – from texts. This study is part of a larger effort to understand the mechanisms of innovation-making through content-aware approaches.

Entity extraction is not a new idea: it is an important sub-task of information extraction and is sometime referred to as named entity extraction and classification (NERC) (Nadeau & Sekine, 2007). The goal of NERC is to identify and classify

* Corresponding author. Tel.: +1 215 895 1459; fax: +1 215 895 2494.
E-mail addresses: ey86@drexel.edu (E. Yan), yz493@drexel.edu (Y. Zhu).

http://dx.doi.org/10.1016/j.joi.2015.04.003
1751-1577/

named entities from large, heterogeneous text corpora. Names, as Kripke (1972) puts it, are "rigid designators" (p. 48). Thus, earlier NERC tasks were largely focused on the extraction of proper names from texts (Thielen, 1995), such as the names of locations, people, and organizations—collectively known as "enamex"; data and time types ("timex") and money and percent types ("numex") were also recognized as entity types for NERC tasks (Nadeau & Sekine, 2007). Recent advances in bioinformatics has also incorporated the identification of biomedical entities, such as genes, compounds, drugs, proteins, and diseases, into the NERC framework (e.g., Bekhuis, 2006; Jensen, Saric, & Bork, 2006; Swanson, Smalheiser, & Torvik, 2006). Meanwhile, we acknowledge the fact that NERC is not restricted to academic research—there are successful commercialized NERC systems for large synchronized language analyses, particularly for defense applications. For instance, the U.S. Defense Advanced Research Projects Agency (DARPA) has allocated more than $100 million between 2003 and 2005 for projects on Automated Speech and Text Exploitation in Multiple Languages (DARPA, 2005). Traditionally, domain specific dictionaries were employed to extract named entities from texts; however, this technique did not scale up with the emergence of new named entities and its performance is impaired by the fuzziness of the natural language (Sekine & Nobata, 2004). Modern statistical methods, on the other hand, are capable of recognizing and disambiguating new named entities, through supervised methods, such as hidden Markov models (HMM; Bikel, Miller, Schwartz, & Weischedel, 1997) and conditional random fields (CRF; Lafferty, McCallum, & Pereira, 2001) or semi- or unsupervised methods, such as bootstrapping (Riloff & Jones, 1999). These methods will be surveyed in the literature review section.

These statistical methods have been applied to extract "enamex", "timex", "numex", and biomedical-related named entities and high precision and recall have been reported (e.g., Collier, Nobata, & Tsujii, 2000; Torii, Hu, Wu, & Liu, 2009; Jiang et al., 2011). However, as Nadeau and Sekine (2007) argued, "[t]he impact of textual genre. . .and domain. . .has been rather neglected in the NERC literature. . .[f]ew studies are specifically devoted to diverse genres and domains" (p.2). Since then, there have been attempts to extend the scope of NERC by extracting entities from scientific literature (e.g., He & Kayaalp, 2008; Prokofyev, Demartini, & Cudré-Mauroux, 2014). Thus, this study is motivated to develop this body of literature by evaluating the performance of several entity extraction methods on a text corpus that contains scientific publications. This textual genre, as a distinctive science communication channel, exhibits its own discourse-related characteristics (Hyland, 2000; Demarest and Sugimoto, in press). Papers in five leading computer science journals are selected as the data set. Several vocabulary- and statistical model-based methods are employed to identify entities from this data set, including two vocabulary-based methods (i.e., a keyword-based and a Wikipedia-based) and three model-based methods (i.e., CRF, CRF with keyword-based dictionary, and CRF with Wikipedia-based dictionary approaches).

Findings from this study will advance the methods of scholarly data mining as well as the application of these methods for content-aware studies of knowledge production and innovation. Conducting content-aware research has the readily apparent advantage of gaining explicit and fine-grained perspectives of how different entities are embedded and related. It will also enhance our understanding of the provenance of knowledge as codified by entities. Results of this research will lay a foundation for these efforts and help inform scientists and scholars for more granular analyses of the history, contemporary landscape, and future trajectories of domains.

2. Related work

This section reviews several types of entity extraction methods, including vocabulary-based, semi- or unsupervised methods, and supervised methods.

2.1. Vocabulary-based methods

Vocabulary-based methods have been employed to identify and disambiguate the concepts of interest, such as title words (e.g., Swanson, 1986), subject headings (e.g., Swanson et al., 2006) and thesaurus dictionaries (e.g., Ding, Chowdhury, & Foo, 2001; Lou & Qiu, 2014). A pioneering study by Swanson (1986) has built off title word co-occurrence relations to detect latent entity relations. Different from finding co-occurrence relations between two directly connected entities, Swanson's approach used two disjoint sets of records and identified a list of terms that co-occurred with both sets. This approach has been empirically tested and it has helped verify some previously overlooked relations such as fish oil and Raynaud's syndrome, magnesium and migraine, somatomedin C and arginine, and even viruses as weapons, according to a review article by Bekhuis (2006). It has been suggested that the use of controlled vocabularies can reduce the ambiguity of the natural language (e.g., Swanson et al., 2006). In biomedical domains, the Medical Subject Headings (MeSH) has been widely used to retrieve medical publications (e.g., Lowe & Barnett, 1994) and to find the relatedness of medical terms (e.g., Nelson, Johnston, & Humphreys, 2001). For instance, Swanson's approach was improved by the use of MeSH terms to enhance its efficiency (Swanson et al., 2006).

Despite the effort of controlled vocabularies such as MeSH to consistently index bio-entities, it was found that they may not fully address the nomenclature problems of synonyms, noun phrases, and acronyms (Morgan, Hirschman, Colosimo, Yeh, & Colombe, 2004; Galvez & de Moya-Anegón, 2012). Thus, more specialized dictionaries and ontologies have been designed and experimented with, serving the goal to discriminate and integrate genes, compounds, drugs, proteins, and diseases in various orthographic forms (e.g., Humphreys, Lindberg, Schoolman, & Barnett, 1998; Ashburner et al., 2000; Jensen et al., 2006; Liu, Hu, Torii, Wu, & Friedman, 2006; Frijters et al., 2008, 2010; Galvez & de Moya-Anegón, 2012). Among these, the Unified Medical Language System (UMLS; Humphreys et al., 1998) and Gene Ontology (GO; Ashburner et al.,