Chapter 22
Techniques for Named Entity Recognition: A Survey

Girish Keshav Palshikar
Tata Research Development and Design Centre, India

ABSTRACT

While building and using a fully semantic understanding of Web contents is a distant goal, named entities (NEs) provide a small, tractable set of elements carrying a well-defined semantics. Generic named entities are names of persons, locations, organizations, phone numbers, and dates, while domain-specific named entities includes names of for example, proteins, enzymes, organisms, genes, cells, et cetera, in the biological domain. An ability to automatically perform named entity recognition (NER) – i.e., identify occurrences of NE in Web contents – can have multiple benefits, such as improving the expressiveness of queries and also improving the quality of the search results. A number of factors make building highly accurate NER a challenging task. Given the importance of NER in semantic processing of text, this chapter presents a detailed survey of NER techniques for English text.

INTRODUCTION

Given the vast amounts of text available on the Web, it is becoming increasingly clear that Internet based tools (e.g., search engines, content creation and management) and applications (e.g., Wikipedia, social networking, blogs) need to understand at least rudimentary semantics of the contents of the Web, which is of course the fundamental motivation for Semantic Web (Shadbolt et al 2006). While semantics is a complex subject and understanding (and using) the complete meaning of a piece of text may well be impossible, it is easy to identify limited types of semantic elements in a text.

Named entities (NE) – like names of persons, organizations, locations and dates, times, phone numbers, amounts, zip codes – are just such basic semantic elements of a text that carry a

DOI: 10.4018/978-1-4666-3604-0.ch022

specific and limited kind of meaning. An ability to automatically perform *named entity recognition* (*NER*) – i.e., identify occurrences of NE in Web contents – can have multiple benefits, such as improving the expressiveness of queries and also improving the quality of the search results. Examples where processing queries containing NE as keywords requires NER: Kawasaki the person and Kawasaki the manufacturing company, Jackson the scientist and Jackson the musician, dates in different formats, identifying that Robert Feynman and Dick Feynman are the same persons, Jobs as a person versus jobs as a common noun, high blood pressure and hypertension as synonymous medical terms. As another example, identifying two successive dates in a text can help compute the duration of some event (e.g., of a project). Lack of NER abilities make representation and execution of queries such as "*Find all European physicists who lived for at least 70 years*" difficult for many of today's search engines. Given the frequent use and relatively well-defined semantics of NE, it is possible to use NER to automatically annotate the occurrences of NE in web contents, which can then be used for improving search and other functions. NE are frequently used as sources (origins) of hyperlinks. Further, since NE occur frequently as part of annotations, notes, comments, bookmarks, hyperlinks etc., NE play an important role in collaborative semantic web applications.

Given the importance of NER in semantic processing of text, this paper presents a detailed (but not necessarily exhaustive) survey of NER techniques. We focus on NER in English text, though there is a considerable work for other languages, which presents complex challenges. We focus mainly on NER for generic NE. However, there is a large amount of work on NER for extracting domain-specific NE. NE in the bio-medical domain are the most well-explored, among the various possible domains.

NER is an important sub-problem in text processing – particularly in *information extraction (IE)* – and is useful in many practical applications in the Semantic Web context. The goal of NER is to identify all occurrences of specific types of *named entities* in the given document collection. NE may be divided into several categories.

- **Generic NE:** Consist of names of persons (PERSON), organizations (ORG), locations (LOCATION), amounts, dates, times, email addresses, URLs, phone numbers etc. Other generic NE include: film title, book title etc. In a richer problem setting (called *fine-gained NER*), the problem is to identify generic NE which are hierarchically organized; e.g., PERSON may be sub-divided into politicians, sports persons, film stars, musicians etc.
- **Domain-specific NE (DSNE):** Consist of, for example, names of proteins, enzymes, organisms, genes, cells etc., in the biological domain. As another example, DSNE in the manufacturing domain are: names of manufacturer, product, brand and attributes of the product (Figure 1).

Figure 1. Example sentences containing occurrences of generic NE

[J. P. Morgan]$_{ORG}$ **strengthens domestic treasury management offering in [Malasia]**$_{LOCATION}$**.**

In a strategic reshuffle at [Bank of America-Merrill Lynch]$_{ORG}$**, [Atul Singh]**$_{PERSON}$ **has taken over as managing director of Global Wealth and Investment Management in [India]**$_{LOCATION}$**.**