A Data-Intensive Approach to Named Entity Recognition Combining Contextual and Intrinsic Indicators

O. Isaac Osesina, University of Arkansas at Little Rock, USA
John R. Talburt, University of Arkansas at Little Rock, USA

ABSTRACT

Over the past decade, huge volumes of valuable information have become available to organizations. However, the existence of a substantial part of the information in unstructured form makes the automated extraction of business intelligence and decision support information from it difficult. By identifying the entities and their roles within unstructured text in a process known as semantic named entity recognition, unstructured text can be made more readily available for traditional business processes. The authors present a novel NER approach that is independent of the text language and subject domain making it applicable within different organizations. It departs from the natural language and machine learning methods in that it leverages the wide availability of huge amounts of data as well as high-performance computing to provide a data-intensive solution. Also, it does not rely on external resources such as dictionaries and gazettes for the language or domain knowledge.

Keywords: Business Intelligence, Data-Intensive Computing, Information Extraction, Semantic Named Entity Recognition (NER), Text Mining, Unstructured Textual Information (UTI)

INTRODUCTION

Innovations in information technology have increased the amount and variety of information accessible by organizations over the past decades. Consumer opinions in online forums and behaviors in online social networking websites are examples of sources of valuable information for businesses intelligence. The use of births, marriages and deaths information from public announcements to update proprietary business databases and specific marketing campaigns are specific examples of how such information can be applied in a business organization. Although effectively analyzing this wide array of information sources can provide businesses competitive advantage in a consumer-centric market environment (Godbole & Roy, 2008), the existence of over 80% of the available information in unstructured form1 (Shilakes & Tylman, 1998) and ineffectiveness of most of the existing automated decision making and

DOI: 10.4018/jbir.2012010104
business intelligence tools (which are primarily designed for structured data) in analyzing unstructured textual information (UTI) is a major challenge (Raghuveer, Jindal, Mokbel, Debnath, & Du, 2007). Several natural language processing (NLP) (Klein & Simmons, 1963) and text mining methods (Church, 1989) have been researched towards solving this problem. Similar to (Wu, et al., 2007), our approach identifies entities as well as their roles within UTI in a process known as semantic named entity recognition (NER) and applies this natural ontology in creating structure for the text. By transforming UTI to structured information, existing business analytics and intelligence extraction processes can be used for processing the information.

Our approach to NER is a language and domain independent data-intensive method. This approach neither utilizes explicit grammar rules nor relies on external resources such as dictionaries and gazettes, making it seamlessly adaptable across different business groups. It derives all the information needed from annotated example documents contained in a knowledgebase. This data-intensive NER approach takes advantage of the ever increasing amount of information available (Cisco Systems, 2007) and the significantly lower cost of high-performance computing resources (Kurzweil, 2005). These two conditions make the capacity to rapidly process this huge volume of information available to businesses with limited resources. Consequently, some difficult problems that were previously solved using complex algorithmic and statistical models are now being approached by more data-intensive solutions that leverages fast computational speeds to extract prior known solutions/information contained in a vast corpora of relevant information (Anderson, 2008). Pioneering this solutions approach are organizations like Google, Yahoo!, Bing and AOL that have access to massive amounts of data. For example, instead of using dictionaries to suggest spelling corrections to search terms, they now use huge amount of information from previous searches, Google translates languages by finding instances of past translations of the same words and phrases (Anderson, 2008). Furthermore, research such as map-reduce by Google (Yang, Dasdan, Hsiao, & Parker, 2007) provides an effective way of utilizing huge volumes of data. This trend towards data-intensive solution can be captured using Peter Norvig, Google’s Research Director’s statement that “all models are wrong, and increasingly you can succeed without them” (Anderson, 2008).

We describe a named entity recognition (NER) approach in which the opportunity for improvement in performance increases along with the volume of information from which knowledge can be extracted. We also discussed its strength and weaknesses using experiments based on a relatively limited knowledgebase.

**NAMED ENTITY RECOGNITION**

NER is the process of locating a word or a phrase that references a particular entity such as a person, organization, place, or event in an unstructured document. For example in the text “XYZ Inc announced on Jan. 1, 2010 that it has acquired controlling interest in ABC Corp,” NER identifies “XYZ Inc” and “ABC Corp” as organizations while identifying “Jan. 1, 2010” as a date. Semantic NER is an important aspect of intelligent information extraction and management research (Jiang, Guan, & Wang, 2006). It determines the semantic role of the entity within the context of the document. In the previous example, semantic NER would recognize “XYZ Inc” is the acquiring organization; “ABC Corp” as the acquired organization while “Jan. 1, 2010” is the acquisition date. The roles and relationships among entities that occur in a family of documents comprise the document ontology.

NLP approach to NER has made very significant progress over the years. Examples of this approach include programming of linguistic rules (Klein & Simmons, 1963) and corpus-based learning approaches (Jelinek, 1990). Different applications used a variety of these approaches along with resources such as WordNet (Fellbaum, 1998), thesauri such as
Related Content

**Improve Intelligence of E-CRM Applications and Customer Behavior in Online Shopping**

**Prediction of Survival and Attrition of Click-and-Mortar Corporations**
Indranil Bose and Anurag Agarwal (2002). *Neural Networks in Business: Techniques and Applications* (pp. 112-123).
[www.igi-global.com/chapter/prediction-survival-attrition-click-mortar/27262?camid=4v1a](www.igi-global.com/chapter/prediction-survival-attrition-click-mortar/27262?camid=4v1a)

**Outlier Detection in Big Data**
[www.igi-global.com/chapter/outlier-detection-in-big-data/107365?camid=4v1a](www.igi-global.com/chapter/outlier-detection-in-big-data/107365?camid=4v1a)
Intelligent Analytics: Integrating Business Intelligence and Web Analytics
Lakshmi S. Iyer and Rajeshwari M. Raman (2011). *International Journal of Business Intelligence Research* (pp. 31-45).
www.igi-global.com/article/intelligent-analytics-integrating-business-intelligence/51557?camid=4v1a