Chapter 3
On Semantic Relation Extraction Over Enterprise Data

Wei Shen
Nankai University, China

Jianyong Wang
Tsinghua University, China & Jiangsu Normal University, China

Ping Luo
Chinese Academy of Sciences, China

Min Wang
Visa Inc., USA

ABSTRACT

Relation extraction from the Web data has attracted a lot of attention recently. However, little work has been done when it comes to the enterprise data regardless of the urgent needs to such work in real applications (e.g., E-discovery). One distinct characteristic of the enterprise data (in comparison with the Web data) is its low redundancy. Previous work on relation extraction from the Web data largely relies on the data's high redundancy level and thus cannot be applied to the enterprise data effectively. This chapter reviews related work on relation extraction and introduces an unsupervised hybrid framework REACTOR for semantic relation extraction over enterprise data. REACTOR combines a statistical method, classification, and clustering to identify various types of relations among entities appearing in the enterprise data automatically. REACTOR was evaluated over a real-world enterprise data set from HP that contains over three million pages and the experimental results show its effectiveness.

INTRODUCTION

Relation extraction is the process of discovering the relationship among two or more entities from a given unstructured data set. It is an important research area not only for information retrieval (Salton & McGill, 1986) but also for Web mining and knowledge base population (Shen, Wang, Luo, & Wang, 2012). The huge amount of valuable information contained in the unstructured text is recorded and

DOI: 10.4018/978-1-5225-5042-6.ch003

transmitted every day in the text form. Turning such information into the understandable and usable form is of high significance and has a lot of real applications.

Traditional relation extraction processes usually require significant human efforts: they need predefined relation names and hand-tagged examples of each named relation as input (Kambhatla, 2004 ; Zelenko, Aone, & Richardella, 2003 ; Giuliano, Lavelli, & Romano, 2006 ; Zhou, Zhang, Ji, & Zhu, 2007 ; Surdeanu & Ciaramita, 2007). Weakly supervised systems for relation extraction such as the bootstrapping systems require much less human involvements, but still require a small set of domain-specific seed instances or seed patterns that have a big impact on the system performance. Furthermore, the seed selection process requires substantial domain knowledge and is usually time consuming (Agichtein & Gravano, 2000 ; Zhu, Nie, Liu, Zhang, & Wen, 2009 ; Brin, 1998 ; Etzioni et al., 2005). Open IE is proposed as a new relation extraction paradigm that can identify various types of relations without predefinition. The goal of open IE systems is to gather a large set of relation facts that can be used for question answering (Banko, Cafarella, Soderl, Broadhead, & Etzioni, 2007 ; Banko & Etzioni, 2008 ; Etzioni et al., 2005 ; Shinyama & Sekine, 2006). Distant supervision approaches require existing knowledge bases (KBs) with which they align an unsupervised text corpus to generate training examples automatically (Hoffmann et al., 2011 ; Madaan et al., 2016 ; Mintz et al., 2009 ; Riedel, Yao, & McCallum, 2010 ; Ritter et al., 2013). Despite that, most relation extraction systems constrain the search for binary relations that are asserted within a single sentence (i.e., single-sentential relations) (Agichtein & Gravano, 2000 ; Hoffmann et al., 2011 ; Mintz et al., 2009 ; Zelenko et al., 2003 ; Brin, 1998 ; Zhu et al., 2009 ; Zhou et al., 2007 ; Hasegawa, Sekine, & Grishman, 2004), while relations between two entities can also be expressed across multiple sentences (i.e., inter-sentential relations). The analysis in Swampillai and Stevenson (2010) shows that inter-sentential relations constitute 28.5% and 9.4% of the total number of relations in MUC6 data set (Grishman & Sundheim, 1996) and ACE03 data set respectively. This places upper bounds on the recall of relation extraction systems that just consider single-sentential relations.

While most work on relation extraction focuses on the Web data, the amount of the enterprise data (including e-mails, internal Web pages, word processing files, and databases) has grown significantly during the past several years for all companies. Numerous real-world entities such as people, organizations, and products are contained in the enterprise data and these entities are connected by various types of relations. To make use of such rich information, it is desirable to build an entity relationship graph that can support efficient retrieval of entities and their relations. A key application of the entity relationship graph is in E-discovery, the process of collecting, preparing, reviewing and producing evidence in the form of Electronically Stored Information (ESI) during litigation (Crowley & Harris, 2007). In this process, lawyers need to find all the people and ESI that are relevant to a legal matter. For example, when a company is alleged to have infringed a patent related to a product, this company is required to disclose all the relevant information. The first question is which employees are closely related to this product. Furthermore, it will be more useful if the method could provide their specific roles to this product, such as product manager, product support, or sales manager. To answer these questions, semantic relation extraction from the enterprise data is an essential step.

However, the existing techniques on relation extraction cannot be applied to the enterprise data directly due to the differences in the data characteristics: the enterprise data has much lower redundancy than the Web data. Figure 1 shows the distribution for the occurrence frequency of entity pairs for the PEOPLE-ORGANIZATION (PEO-ORG) domain in the enterprise data set used in the experiments. It shows that more than 90% of the entity pairs occur less than four times, about two thirds of the entity pairs only occur once in the entire data set and the average occurrence frequency of all the entity pairs is