Expert Systems with Applications 39 (2012) 2690–2699

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

A robust web personal name information extraction system

Ying Chen [a,c,*], Sophia Yat Mei Lee [b,c], Chu-Ren Huang [c]

[a] College of Information and Electrical Engineering, China Agricultural University, PR China
[b] Language Centre, Hong Kong Baptist University, Hong Kong
[c] Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong

ARTICLE INFO

Keywords:
Information extraction
Named entity disambiguation
Attribute extraction
Relation extraction

ABSTRACT

Personal information extraction, which extracts the persons in question and their related information (such as biographical information and occupation) from web, is an important component to construct social network (a kind of semantic web). For this practical task, two important issues are to be discussed: personal named entity ambiguity and the extraction of personal information for a specific person. For personal named entity ambiguity, which is a common phenomenon in the fast growing web resource, we propose a robust system which extracts lightweight features with a totally unsupervised approach from broad resources. The experiments show that these lightweight features not only improve the performances, but also increase the robustness of a disambiguation system. To extract the information of the focus person, an integrated system is introduced, which is able to effectively re-use and combine current well-developed tools for web data, and at the same time, to identify the expression properties of web data. We show that our flexible extraction system achieves state-of-the-art performances, especially the high precision, which is very important for real applications.

1. Introduction

Social network is a kind of semantic web which collects different kinds of information of people. It plays an important role in the development of a new generation of web. However, it is not easy to construct social network not only because relations among persons are complicated but also due to the noisy information in web. For example, the name "Clinton" may refer to Bill Clinton or Hillary Clinton. Besides their relationship, both Bill and Hillary occupy different positions from time to time. While Bill Clinton served as the president of the US from 1993 to 2001, apart from being the First Lady, Hillary Clinton was also the Founding Chair of the Save America's Treasures program, the head of the White House Millennium Council, and so on; While Hillary Clinton has been the Secretary of State from 2008, Bill Clinton is still active in the political stage, such as initiating the trip to North Korea. In this paper, we discuss two practical issues related to the construction of social network: personal named entity ambiguity and the information extraction for a specific person. We then develop an information extraction (IE) system, which provides a solution to the two problems.

Information extraction (IE), if it refers to a broad area, ranges from named-entity detection and recognition (NER) to attribute (information) extractions for the specific named-entity (AE). AE is sometimes considered as a limited concept of IE. One big problem, which web NER needs to deal with, is named-entity ambiguity, i.e. two different persons share the same name. Another challenge of developing a named-entity disambiguation system is the maintenance of the system's robustness. For example, a personal name disambiguation system needs to maintain the performance for different personal names in web even when there are varying ambiguities of those personal names. Furthermore, given a named-entity, its complete or comprehensive information often needs to be extracted from the related web pages; however, this information usually comes from many different sources with different formats. For example, the information of "affiliation" and "email" of a person is different in terms of expressions so that they need different extraction approaches. Hence, most of the existing homogeneous AE systems, no matter statistical or rule-based, often cannot perform effectively for web AE. In this paper, we focus on the most popular named-entity, i.e. personal named-entity, and attempt to tackle two problems with approaches differing from the previous work, i.e. how to improve the robustness of a personal name disambiguation system with lightweight features, and how to effectively integrate the heterogeneous AE approaches into web personal AE.

For personal name disambiguation, many systems, which participated in the WePS[1] bakeoff, use a combination of feature

* Corresponding author at: College of Information and Electrical Engineering, China Agricultural University, PR China. Tel.: +86 1062738792.
E-mail addresses: chenying3176@gmail.com (Y. Chen), sophiaym@gmail.com (S.Y. Mei Lee), churenhuang@gmail.com (C.-R. Huang).

[1] http://nlp.uned.es/weps/

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.08.125

extractions followed by clustering to disambiguate web personal names. The features employed include simple tokens, base syntactic chunks, named-entities, dependency parses, semantic role labels, etc. For the most parts, these features are extracted using off-the-shelf NLP systems designed to annotate or extract the relevant information. Unfortunately, web data are quite diverse and differ fundamentally from the news-oriented sources that have traditionally been the source of training material for those NLP systems used in feature extraction. This leads to a severe degradation in the performance of the individual feature extractors and the subsequent clustering algorithms. In this paper, we attempt to overcome these difficulties using lightweight features from web-derived resources.

For web personal AE, a crucial problem is that most of the current AE systems are developed only for news articles, and their adoption to web data is not an easy task (Vilain, Su, & Lubar, 2007). Although some previous systems (Culotta, Bekkerman, & McCallum, 2004; Lan, Zhang, Lu, & Su, 2009; Watanabe, Bollegala, Matsuo, & Ishizuka, 2009) have tried to combine different AE approaches for web data, nevertheless, few of them have explored how to effectively utilize or integrate different AE tools for web data. Instead of presenting a new AE solution for web data, the goal of this paper aims at providing a flexible framework which can effectively reuse and integrate the existing well-developed heterogeneous AE technologies according to different text expression formats in web data.

Our IE system, including a disambiguation system and an AE system, participated in the WePS 2009 bakeoff, and achieved the best performances. Web People Search (WePS) (Artiles et al., 2009; Sekine & Artiles, 2009) provides a forum for a standard evaluation, which focuses on IE for web persons. It includes two tasks: clustering and attribute extraction (AE). The clustering task, which is also called personal name disambiguation, groups the web pages according to whether the given personal name occurring in that web page refers to the same person in reality. Attribute extraction extracts certain personal information of a target person in a web page.

The remainder of this paper is organized as follows. Section 2 addresses the related work of both personal name disambiguation and AE. Section 3 introduces our personal name disambiguation approach, and Section 4 describes our flexible personal AE framework. Section 5 presents the experiments, which run on the WePS 2009 corpus, and gives further analysis. Finally, some conclusions are drawn in Section 6.

2. Previous work

For a given set of web pages containing a focus personal name, a cascaded structure usually is adopted to extract personal information. Firstly, a personal name disambiguation system is used to group those web pages according to which persons they refer to. Then, for each person in interest, an AE system is applied to extract personal information from its corresponding group of web pages. In the following section, we describe the related work for each component.

2.1. Personal name disambiguation

Named-entity co-reference, a component of NER, can be divided into within-document co-reference and cross-document co-reference. One important difference between within-document co-reference and cross-document co-reference is the different phenomena of named-entity ambiguity. Mentions with the same string in a document are very likely to refer to the same entity, whereas it is not true for the ones among different documents. Therefore, it is essential for web NER, which comprises cross-document co-reference, to handle this kind of named-entity ambiguity problem.

Personal name ambiguity is so common in web that most previous disambiguation systems choose to work on personal name disambiguation (Bagga & Baldwin, 1998; Gooi & Allan, 2004; Li, Xin, & Dan Roth, 2004; Niu, Li, & Srihari, 2004). Due to the lack of a large labeled corpus and the varying ambiguity of different personal names, most previous personal disambiguation systems use unsupervised clustering as the basic approach, despite the different features used to create the similarity space. There are two kinds of features: document-level features and global features. Document-level features refer to the information extracted from the given corpus, such as tokens in a given web page (Bagga & Baldwin, 1998; Gooi & Allan, 2004; Pedersen, Ted, Purandare, & Kulkarni, 2005), biographical information (Mann, 2006), and bigrams (Pedersen & Kulkarni, 2007). Global features refer to the information derived beyond the given corpus, such as the information from an extra corpus (Kalmar & Blume, 2007; Mann, 2006; Pedersen & Kulkarni, 2007), and online information (Rao, Garera, Yarowsky, & David, 2007). Global feature extraction is comparatively less popular than document-level feature extraction. Some previous works using global features are given as follows.

Mann (2006), Kalmar and Blume (2007) find that the given corpus is often not large enough to learn the realistic probabilities or weights for those document-level features, therefore they added more data into the given corpus in order to give a realistic probability for a token (Mann, 2006) or a realistic frequency for a proper name (Kalmar & Blume, 2007). In addition, Pedersen and Kulkarni (2007) use an extra corpus to filter out some un-collocation bigrams in the given corpus, and then create a token-based representation for the bigram feature with the help of scores that measure whether a bigram is a collocation.

Besides the given corpus, there is a large amount of online information about the focus person. Since most of the given corpora include only a small part of documents containing the ambiguous personal name, such as the top 100 web pages in the WePS 2009 corpus, it is sometimes hard to make a co-reference decision, even for an annotator. To alleviate this shortcoming, for each ambiguous personal name, Rao et al. (2007) retrieve 1,000 snippets through the query of the focus personal name in the Google search engine, and merge these low-noise snippets into the training or test data. Each snippet is treated as a document, and can be served as a bridge. For example, if two web pages referring to the same named-entity have not enough shared information in the given corpus, the additional snippet-based information can sometimes provide a bridge to connect them.

2.2. Attribute extraction

Although AE is an old topic, it still poses a great challenge, especially for web data. In general, AE can be divided into two steps: the detection of attribute value candidates and attribute value assignment. The detection of attribute value candidates can be considered as a specific case of NER task because most attribute values are named-entities. The specific NER extracts all possible candidates regarding the attribute on question. Attribute value assignment is often done through relation detection and recognition (RDR): for each attribute value candidate detected by NER, if a specific relation, which can be considered as a rename of a specific attribute, exists between the focus named-entity and the given candidate, it is assigned as a value of the specific attribute. To effectively tackle AE, both specific NER and RDR are required to have a good performance.

For NER, the naïve approach is rule-based, but its big disadvantage is the difficulty of rule-design (Feldman, 2002) or rule-learning (Etzioni & Cafarella, 2005). Statistical NER technology (Bikel,