An Internet-based Method for Verification of Extracted Proper Names

Angelo Dalli

NLP Research Group
Department of Computer Science
University of Sheffield
angelo@dcs.shef.ac.uk

Abstract. Identification and extraction of proper names from Internet-based sources currently suffers from a lack of verification methods that check the validity of these extracted names. A language-independent method for assigning probabilities to extracted proper names using frequency data harvested from the Internet is presented. Verification mechanisms are built on top of this technique to exclude misidentified proper names automatically.

1 Introduction

The task of automatically identifying proper names in text is often aided through gazetteer lists obtained from telephone directories, government records, and additional sources together with additional heuristics. Although traditional publication sources provide acceptably large gazetteer lists for a single language, it may prove difficult to integrate different data sources to produce a comprehensive multi-lingual gazetteer list.

An innovative approach to gathering a large set of proper names was adopted by having a small custom-built information extraction system analyse a large corpus of multi-lingual text crawled over the Internet. Simple capitalisation rules together with the presence of various personal titles such as Mr., Ms., and so on were used to identify likely proper names in the texts. Additional hints were provided by the presence of anaphora in the same sentence or the following sentence as the suspected proper name. A manually compiled list of anaphora was used to detect their presence. The gender of every title and anaphora was manually noted and this information was used to keep a count of the number of male or female titles and anaphors associated with a particular name. This enabled the list of names to be organised by gender. Although this is not the most sophisticated form of multi-lingual anaphora detection and resolution, acceptable results were obtained since the task was only to assign a rough probability to suspect words [1–7].

One of the main problems encountered in this approach were the numerous nicknames used in online chat rooms, forums and other communications that were often incorrectly identified as proper names. This problem was solved partially by removing names that had unusual punctuation marks or numeric digits.