Language Independent First and Last Name Identification in Person Names

Octavian Popescu[1] and Bernardo Magnini[1]

[1] FBK-Trento, Italy
{popescu, magnini}@fbk.eu

Abstract. In this paper we address the problem of first name and last name identification in a news collection. The approach presented is based on corpus investigation and is language independent. At the core of the system there is a name classifier based on the values of different parameters. In its most general form, the name category identification is not an easy task. The hardest problems are raised by ambiguous tokens – those that can be either a first or a last name and/or by tokens with just one occurrence. However, the system is able to predict the name category with high accuracy. The experiments have been run on an Italian newspaper and the evaluation has been carried on I-CAB.

1 Introduction

Knowing whether a token composing the name of a person refers either to her/his first or last name is an important task in several respects. It is probably one of the first things a reader would like to know about a person, especially if she/he has no native intuitions. The name category (first vs. last) is also important for further processing of textual information. It plays an important role in enhancing the overall accuracy of a cross document coreference system (Popescu&Magnini, 2007). Many name mentions consist of only one token, but they definitely stand for two-token names; knowing the category of the token that is missing may give important clues about the person that carries that name.

The task consists in determining for each name occurrence in a large corpus the name category of each individual token composing it. For example, "*George W. Bush*" should be analyzed like "*George*<first name> *W.*<first name> *Bush*<last name>". In its most general form the name category identification is not an easy task. The hardest problems are raised by ambiguous tokens – those that can be either first or last names and/or by tokens with just one occurrence.

In this paper we address the problem of first, last name identification in a news collection. While relying on gazetteers or name dictionaries seems to be an easy way out, we show that this is not enough. The approach we are going to present is based on