An Unsupervised Language Independent Method of Name Discrimination Using Second Order Co-Occurrence Features

Ted Pedersen[1], Anagha Kulkarni[1], Roxana Angheluta[2],
Zornitsa Kozareva[3], and Thamar Solorio[4]

[1] University of Minnesota, Duluth, USA
[2] Katholieke Universiteit Leuven, Belgium
[3] University of Alicante, Spain
[4] University of Texas at El Paso, USA

Abstract. Previous work by Pedersen, Purandare and Kulkarni (2005) has resulted in an unsupervised method of name discrimination that represents the context in which an ambiguous name occurs using second order co–occurrence features. These contexts are then clustered in order to identify which are associated with different underlying named entities. It also extracts descriptive and discriminating bigrams from each of the discovered clusters in order to serve as identifying labels. These methods have been shown to perform well with English text, although we believe them to be language independent since they rely on lexical features and use no syntactic features or external knowledge sources. In this paper we apply this methodology in exactly the same way to Bulgarian, English, Romanian, and Spanish corpora. We find that it attains discrimination accuracy that is consistently well above that of a majority classifier, thus providing support for the hypothesis that the method is language independent.

1 Introduction

Purandare and Pedersen (e.g., [9], [10]) previously developed an unsupervised method of word sense discrimination that has also been applied to name discrimination by Pedersen, Purandare, and Kulkarni [8]. This method is characterized by a reliance on lexical features, and avoids the use of syntactic or other language dependent information. This is by design, since the method is intended to port easily and effectively to a range of languages. However, all previous results with this method have been reported for English only.

In this paper, we evaluate the hypothesis that this method of name discrimination is language independent by applying it to name discrimination problems in Bulgarian, Romanian, and Spanish, as well as in English.

Ambiguity in names of people, places and organizations is an increasingly common problem as online sources of information grow in size and coverage. For example, Web searches for names frequently locate different entities that share