Name Discrimination by Clustering Similar Contexts

Ted Pedersen[1], Amruta Purandare[2], and Anagha Kulkarni[1]

[1] University of Minnesota, Duluth, MN 55812, USA
[2] University of Pittsburgh, Pittsburgh, PA 15260, USA
http://senseclusters.sourceforge.net

Abstract. It is relatively common for different people or organizations to share the same name. Given the increasing amount of information available online, this results in the ever growing possibility of finding misleading or incorrect information due to confusion caused by an ambiguous name. This paper presents an unsupervised approach that resolves name ambiguity by clustering the instances of a given name into groups, each of which is associated with a distinct underlying entity. The features we employ to represent the context of an ambiguous name are statistically significant bigrams that occur in the same context as the ambiguous name. From these features we create a co–occurrence matrix where the rows and columns represent the first and second words in bigrams, and the cells contain their log–likelihood scores. Then we represent each of the contexts in which an ambiguous name appears with a second order context vector. This is created by taking the average of the vectors from the co–occurrence matrix associated with the words that make up each context. This creates a high dimensional "instance by word" matrix that is reduced to its most significant dimensions by Singular Value Decomposition (SVD). The different "meanings" of a name are discriminated by clustering these second order context vectors with the method of Repeated Bisections. We evaluate this approach by conflating pairs of names found in a large corpus of text to create ambiguous pseudo-names. We find that our method is significantly more accurate than the majority classifier, and that the best results are obtained by having a small amount of local context to represent the instance, along with a larger amount of context for identifying features, or vice versa.

1 Introduction

The problem of name ambiguity exists in many forms. It is common for different people to share the same name. For example, there is a George Miller who is a prominent Professor of Psychology, another who is a Congressman from California, and two more who are film directors from Australia. Locations may have the same name. For example, Duluth is a city in Minnesota and also a city in Georgia. The acronyms associated with organizations may also be ambiguous. UMD can refer to the University of Michigan – Dearborn, the University of Minnesota, Duluth or the University of Maryland .