Reported at the 57th Session of the International Statistics Institute

Implicit translation of Wikipediae via Random Graph Embeddings

D.J. Marchette, E. Hohman, C.E. Priebe
Naval Surface Warfare Center
Johns Hopkins University
dmarchette@gmail.com

The multilingual Wikipediae provide a good testbed for developing methods for the analysis of text and the fusion of text and graph information. In this work we focus on the problem of implicit translation: given a Wikipedia in one language L_1 and another in L_2, with some known associations of articles in one to the other, can one determine further associations, without an explicit translation dictionary; for example, determining that the Afrikaans article titled "Sterrekunde" should be matched with the one titled "Astronomie" in Dutch or "Astronomy" in English. To perform the matching, we define a novel random projection which allows us to project the Wikipediae into the same space, with similarity in this space providing the association. We explore the performance of this method for pairing Afrikaans and Dutch articles, and Afrikaans and articles written in several Bantu languages. The English associations provided by the Wikipediae allow an objective performance evaluation. The random embedding method is designed to be applicable to very large graphs, and we demonstrate it on both the small Bantu graphs (about 100 articles each) and the large graphs in Afrikaans (10K articles) and Dutch (490K articles). Figure 1 depicts the distances between matched articles (pairs of Afrikaans and Dutch articles with the same title) compared with the distances between unmatched articles, for a 10-dimensional random projection. The figure shows that matched documents are closer to each other than they are to unmatched documents.

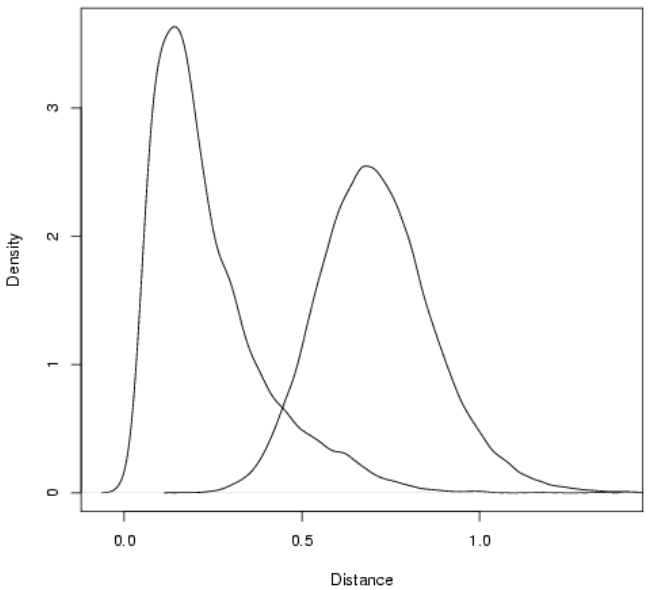

Figure 1: Distances between matched pairs (left curve) and unmatched pairs (right curve) for the Afrikaans and Dutch Wikipediæ.