Chapter XI

A Kernel Canonical Correlation Analysis for Learning the Semantics of Text

Blaž Fortuna, Jožef Stefan Institute, Slovenia

Nello Cristianini, University of Bristol, UK

John Shawe-Taylor, University of Southampton, UK

Abstract

We present a general method using kernel canonical correlation analysis (KCCA) to learn a semantic of text from an aligned multilingual collection of text documents. The semantic space provides a language-independent representation of text and enables a comparison between the text documents from different languages. In experiments, we apply the KCCA to the cross-lingual retrieval of text documents, where the text query is written in only one language, and to cross-lingual text categorization, where we trained a cross-lingual classifier.
Introduction

The ideas of using statistical techniques from communication theory to translate text from one natural language to another have been around for many years. Literally, computational text analysis is as old as computers (which were introduced as part of UK decryption efforts in World War II).

A pioneer of information theory, Warren Weaver suggested automatic approaches to cross-language analysis as early as 1949, but efforts in this direction were soon abandoned for various technical and theoretical reasons (Weaver, 1955). The idea of automatically detecting relations between languages has not been abandoned, however. Today, state-of-the-art methods in automatic translation rely on statistical modeling.

In the mid-'30s a statistical technique had been introduced to find correlations between two sets of vectors: canonical correlation analysis (CCA). Various computational and conceptual limitations made a statistical analysis of bilingual text based on CCA impractical (Hoteling, 1936). For one, the vector space model of text documents became popular in information retrieval (IR) only in the 1970s. The recent combination of CCA with kernel methods (Bach & Jordan, 2002; Lai & Fyfe, 2000; Shawe-Taylor & Cristianini, 2004) paved the way to a number of statistical, computational, and conceptual extensions. Various kernels were used to represent semantic relations, dual representation to speed up computations, and regularization-theory ideas to avoid discovering spurious correlations.

We present in this chapter a method based on kernel CCA to extract correlations between documents written in different languages with the same content. Given a paired bilingual corpus (a set of pairs of documents, each pair being formed by two versions of the same text in two different languages), this method defines two embedding spaces for the documents of the corpus, one for each language, and an obvious one-to-one correspondence between points in the two spaces. KCCA then finds projections in the two embedding spaces for which the resulting projected values are highly correlated. In other words, it looks for particular combinations of words that appear to have the same co-occurrence patterns in the two languages. Our hypothesis is that finding such correlations across a paired cross-lingual corpus will locate the underlying semantics since we assume that the two languages are conditionally independent, or that the only thing they have in common is their meaning.

The directions would carry information about the concepts that stood behind the process of the generation of the text and, although expressed differently in different languages, are, nevertheless, semantically equivalent. A preliminary version of this study appeared in Vinokourov, Shawe-Taylor, and Cristianini (2002).

Other methods exist to analyze statistically relations within bilingual text, such as cross-lingual latent semantic indexing (LSI; Littman, Dumais, & Landauer, 1998). LSI (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) has been used to extract information about the co-occurrence of terms in the same documents, an indicator of semantic relations, and this is achieved by singular value decomposition (SVD) of the term-document matrix. The LSI method has been adapted to deal with the important problem of cross-language retrieval, where a query in a language is used to retrieve documents in a different language. Using a paired corpus, after merging each pair into a single document, we can interpret frequent