Multilingual Text Mining

Peter A. Chew
Sandia National Laboratories, USA

INTRODUCTION

The principles of text mining are fundamental to technology in everyday use. The world wide web (WWW) has in many senses driven research in text mining, and with the growth of the WWW, applications of text mining (like search engines) have by now become commonplace. In a way that was not true even less than a decade ago, it is taken for granted that the 'needle in the haystack' can quickly be found among large volumes of text. In most cases, however, users still expect search engines to return results in the same language as that of the query, perhaps the language best understood by the user, or the language in which text is most likely to be available.

The distribution of languages on the WWW does not match the distribution of languages spoken in general by the world's population. For example, while English is spoken by under 10% of the world's population (Gordon 2005), it is still predominant on the WWW, accounting for perhaps two-thirds of documents. There are variety of possible reasons for this disparity, including technological inequities between different parts of the world and the fact that the WWW had its genesis in an English-speaking country. Whatever the cause for the dominance of English, the fact that two-thirds of the WWW is in one language is, in all likelihood, a major reason that the concept of *multilingual* text mining is still relatively new. Until recently, there simply has not been a significant and widespread need for multilingual text mining.

A number of recent developments have begun to change the situation, however. Perhaps these developments can be grouped under the general rubric of 'globalization'. They include the increasing adoption, use, and popularization of the WWW in non-English-speaking societies; the trend towards political integration of diverse linguistic communities (highly evident, for example, in the European Union); and a growing interest in understanding social, technological and political developments in other parts of the world. All these developments contribute to a greater demand for multilingual text processing – essentially, methods for handling, managing, and comparing documents in multiple languages, some of which may not even be known to the end user.

BACKGROUND

A very general and widely-used model for text mining is the vector space model; for a detailed introduction, the reader should consult an information retrieval textbook such as Baeza-Yates & Ribeiro-Neto (1999). Essentially, all variants of the vector space model are based on the insight that documents (or, more generally, chunks of text) can also be thought of as vectors (or columns of a matrix) in which the rows correspond to terms that occur in those documents. The vectors/matrices can be populated by numerical values corresponding to the frequencies of occurrence of particular terms in particular documents, or, more commonly, to *weighted* frequencies. A variety of weighting schemes are employed; an overview of some of these is given in Dumais (1991). A common practice, before processing, is to eliminate rows in the vectors/matrices corresponding to 'stopwords' (Luhn, 1957) – in other words, to ignore from consideration any terms which are considered to be so common that they contribute little to discriminating between documents. At its heart, the vector space model effectively makes the assumption that the meaning of text is an aggregation of the meaning of all the words in the text, and that meaning can be represented in a multidimensional 'concept space'. Two documents which are similar in meaning will contain many of the same terms, and hence have similar vectors. Furthermore, 'similarity' can be quantified using this model; the similarity of two documents in the vector space is the cosine between the vectors for the documents. Document vectors in the vector space model can also be used for supervised predictive mining; an example is in Pang et al. (2002), where document vectors are used to classify movie reviews into 'positive' versus 'negative'.

A variant on the vector space model commonly used in text mining is Latent Semantic Analysis (LSA) (Deerwester et al., 1990). This approach takes advantage of the higher-order structure in the association of terms with documents by applying singular value decomposition (SVD) to the initial term-by-document matrix. SVD is a method in multilinear algebra for decomposition of a matrix into its principal components, and is used to find the best lower-rank approximation to the original matrix. In text mining, SVD can be used to '[represent] both terms and documents as vectors in a space of choosable dimensionality' (Deerwester et al., 1990: 395). The principal contribution of LSA is that it deals with the inherently noisy characteristics of text by mapping terms to more nebulous 'concepts' without abandoning statistical principles to do so; a thesaurus is not required to find terms of similar meaning, as terms of similar meaning should in theory be identifiable by the statistics of their distributions. When applied to practical text mining problems, the results of LSA analyses have often been shown to be as good as, or better than, simpler vector space approaches (Deerwester et al., 1990). Perhaps the most significant advantage of LSA, however, is that the representation of documents is generally much more economical than under approaches in which each term corresponds to the 'dimension' of a vector. In typical real-world text mining problems, document sets may contain thousands of distinct terms, meaning that, with simple vector-space approaches, document vectors must contain thousands of entries (although the vectors are usually sparse). With LSA, on the other hand, documents are often represented by vectors containing on the order of 200-300 values. When it comes to computing cosines between vectors or training predictive data mining models, for example, this can save a significant amount of computational horsepower.

MAIN FOCUS

The basic problem of multilingual text mining (also known as cross-language information retrieval) is that of representing text from different languages in a single, coherent conceptual space. If this can be achieved, then, at least in theory, the groundwork is laid for solving problems such as the following:

- Computation of the similarity of documents in different languages
- Prediction of other variables (such as positive or negative sentiment) from the text, regardless of the text's language
- Clustering of documents in multiple languages by topic, regardless of the documents' languages

From a multilingual perspective, the vector space model in general has many practical features which recommend it. First, the process of tokenization – computing which terms occur in which documents – is very general, from a linguistic point of view. While linguistically the concept of the 'word' may carry different connotations from language to language, it is computationally straightforward to define the 'term' (which at least in English often corresponds to the word) in a way which can easily be applied to virtually all languages. The task is made easier by the fact that a computational infrastructure for this already exists: both Unicode, itself a product of the 'globalization' of computing, and regular expressions, facilitate tokenization in multiple languages. Essentially, we can say that a term is any portion of text bounded at its left and right edges by 'non-word' characters (the latter being pre-defined in the regular expressions framework for virtually all Unicode code pages). Thus, Unicode allows this definition to be applied equally for languages in Roman and non-Roman script; for languages with left-to-right or right-to-left script; and so on. Of the major languages of the WWW, Chinese is perhaps the only one that presents a challenge to the general definition, as 'non-word' characters often occur only at sentence boundaries in Chinese. Yet this is by no means an insurmountable problem; while Chinese characters do not exactly correspond to English words, a reasonable approach with Chinese, representing only a minor deviation from the general rule proposed above, would be to treat each *character* as a term, since Chinese characters in general each have a meaning of their own.

The simplest vector space model, in which the rows of the vectors and matrices correspond one-to-one to distinct terms, is inappropriate for multilingual text mining, however. The reason for this is that there is usually comparatively little overlap between languages with respect to the terms that occur in those languages. Even where it exists, some of the overlap is attributable to 'faux amis' – words which are homographic across languages but have different meanings, such as