Disentangling from Babylonian Confusion – Unsupervised Language Identification

Chris Biemann, Sven Teresniak

Leipzig University
Computer Science Institute, NLP Dept.
Augustusplatz 10/11
04109 Leipzig, Germany
biem@informatik.uni-leipzig.de, knorke@zehnvierzig.org

Abstract: This work presents an unsupervised solution to language identification. The method sorts multilingual text corpora on the basis of sentences into the different languages that are contained and makes no assumptions on the number or size of the monolingual fractions. Evaluation on 7-lingual corpora and bilingual corpora show that the quality of classification is comparable to supervised approaches and works almost error-free from 100 sentences per language on.

1 Introduction

With a growing need for text corpora of various languages in mind, we address the question of how to build monolingual corpora from multilingual sources without providing training data for the different languages.

According to [Pantel et al. 2004], shallow methods of text processing can yield comparable results to deep methods when allowing them to operate on large corpora. The larger the corpus, however, the more difficult it is to ensure sufficient quality. Most approaches in Computational Linguistics work on monolingual resources and will be disturbed or even fail if a considerable amount of 'dirt' (sublanguages or different languages) are contained. Viewing the Internet as the world's largest text corpus, it is difficult to extract monolingual parts of it, even when restricting downloading to country domains or some domain servers.

While some languages can be identified easily due to their unique coding ranges in ASCII or UNICODE (like Greek, Thai, Korean, Japanese and Chinese), the main difficulty arises in the discrimination of languages that use the same coding and some common words, as most of the European languages do.

In the past, a variety of tools have been developed to classify text with respect to its language. The most popular system, the *TextCat Language Guesser* as described in [Cavnar & Trenkle 1994], makes use of the language-specific letter N-gram distribution and can determine 69 different natural languages. According to [Dunning 1994], letter trigrams can identify the language almost error-freely from a text-length of 500 bytes on. Other language identification approaches use short words and