Demonstration Abstract:

Written Language Identification Using Word Fragments of Variable Length

Alexander Troussov, Mayo Takeuchi, D.J.McCloskey
IBM LanguageWare Group
IBM Dublin Software Lab, Airways Ind. Est., Cloghran, Dublin 17, Ireland
Contact author: Alexander Troussov ATrousso@ie.ibm.com

In multi-lingual environments, identifying the language of a piece of text is usually a prerequisite for subsequent processing. Language identification systems have been reported to have nearly perfect performance. However, development of efficient solutions still remains an important practical problem for domains with severe constraints on the size of the analyzed texts and on computational resources. The language identification system "Fragma" presented in this paper provides a uniform and computationally efficient implementation of the main ideas behind two major language identification schemes – Method of Words and character N-grams.

During the language modeling stage, a small number of the most salient features was shortlisted for each individual language (for most alphabetical languages the number of features measured in hundreds). Features might represent alphabet, phonetics, roots, derivational suffixes. Special attention was paid to features important for discourse, like inflectional suffixes and function words. Negative features (for example, presence of Cyrillic letters could be considered to be a "negative" feature of English), were extensively used to improve both recall and precision of the system when analyzed texts might be written in languages that are not supported. All word fragments are compiled into one amalgamated finite-state system dictionary.

"Fragma" works on texts that are encoded in Unicode. During the text modeling stage, input text is segmented into tokens by a parser that exploits ICU (International Components for Unicode) properties of the characters. An additional regular expressions mechanism is used both for chunking and for detecting non-lexical items such as www-addresses or meta-tags so that they do not interfere with language identification. Parsing is crosslinguistic regarding language or script, including languages that use ideographics and/or without explicit word boundaries. The system searches the lexical tokens and detects the presence of features stored in the system dictionary. In order to discriminate between languages using Arabic script, we additionally use partial morphological analysis (morphosyntactic compatibility of prefixes and postfixes found in the same surface form).

"Fragma" was implemented in order to categorize most languages that are of commercial importance. The Java version processes 20 GB/hour on an average PC. With a data repository of 1.2 MB for 25 languages, the system achieves 95% accuracy on texts ranging in length between 76 and 100 characters ("soft errors" like Norwegian mistaken for Danish were counted as errors). The system was tested for non-discourse texts such as query logs of computer systems. In addition to language identification, the system was tested as a generic text categorizer (for example, to identify biomedical texts).

A demonstration of the system will be available at the conference.