A web-based tool for online language recognition, text segmentation and name recognition of texts

by Lilian Karlsson
Department of Computer Science
University of Gävle
801 76 Gävle, Sweden

Abstract: This proposal is about PALaTe-LTN (PALaTe-Language identification, Text segmentation, Name recognition), a web-based tool for online language recognition, text segmentation and name recognition of texts, recently developed in the PALaTe project. As time for teacher/student-interaction is reduced, students have to perform more self-studies than before. This calls for effective modularization of learning process, which help students to prepare for courses. It also requires the provision of additional background information as well as documentation after the students have completed their courses. Since large textual resources are often objects of study in Language and Humanities education, the PALaTe-LTN considers the ability to navigate and analyze in large text materials. The tool has the goals to break down an input text to different segments and to do some language recognition and name recognition.

Keywords: language recognition, text segmentation, name recognition, self-studies.

Personalized Access to Distributed Learning Repositories (PADLR) develops, tests and evaluates a new infrastructure and new tools for individual use of learning materials. The PADLR project is a collaboration between Stanford Learning Lab (Stanford University), Learning Lab Lower Saxony (Universities of Hannover, Braunschweig and Karlsruhe) and Swedish Learning Lab. The Swedish Learning Lab elaborates test tools for one of the PADLR subprojects - PALaTe (Personalized Access to Large Text Archives). Since large textual resources are often objects of study in Language and Humanities education, the main tool of the PALaTe project - PALaTe-LTN considers the ability to navigate and analyse long text materials.

PALaTe-LTN consists of two large modules and is available for use on the Web. The first module identifies natural language in input digital text and returns as a result, which languages are occurring in the text. It is able to support an unlimited number of languages, but currently there are only a few languages loaded into the system. The second module splits an input text in digital form into several text segments. Both modules are available on a Linux server and use a mutual interface on the Web. A user sends input text via a simple graphic interface for analysis and receives the result on the same interface. A third module is under development; namely modules for name recognition.

Students, teachers and researchers in humanitarian science can use PALaTe-LTN; especially when analysis of large text material is needed. The probability to get more correct results is increased if the system receives larger text files.

PALaTe-LTN’s main design is a client-server model. The server part contains of striper of HTML/SGML tags, modules for language recognition and text segmentation and a simple file database. The client part consists of a number of scripts that interact with the modules and with the user through a web browser. Files uploaded should be encoded with the UTF-8 Unicode encoding.

PALaTe-LTN has been developed using the C programming language and is activated via several CGI scripts. Remaining scripts are written in Java and Perl, which are activated from the applications interface written in HTML and Java script. An input file for analysis must have HTML/SGML or text format. The result of the analysis is presented in RDF (Resource Description Framework) format, stored on the server and displayed on the Web interface. PALaTe-LTN supports Unicode and is available on a server (http://130.238.162.205:8080).

I will discuss some of the practical and theoretical considerations, which have influenced the design of PALaTe-LTN, and also demonstrate it.