atlas | project

Embedding Machine Translation in ATLAS Content Management System

EU CIP-ICT Policy Support Programme Funding call identification
CIP-ICT-PSP-2009-3 Theme 3 Multilingual Web
Project ID number : 250467
http://www.atlasproject.eu

List of partners
Tetracom Interractive Solutions Ltd., Bulgaria (coordinator)
Institute for Bulgarian Language at the Bulgarian Academy of Sciences, Bulgaria
Institute of Technology and Development, Bulgaria
University of Hamburg - Research Group "Computerphilology", Germany
German Research Center for Artificial Intelligence, Germany
Atlantis Consulting SA, Greece
Institute of Computer Science of the Polish Academy of Sciences, Poland
Alexandru Ioan Cuza University of Iasi, Romania

Project duration: March 2010 — February 2013

Summary

The project aims to adjust and integrate several existing software components, assembling a platform for multilingual web content management called ATLAS, and a visualization layer called i-Publisher, which adds to the platform a powerful web-based point-and-click tool for building, reusing and managing multilingual content-driven web sites. ATLAS makes use of state-of-the art text technology methods in order to extract, translate information and cluster documents according to a given hierarchy. With the current available technology it is not possible to provide a translation system which is domain- and language variation independent and works for a couple of heterogeneous language pairs. Thus our approach envisages a system of user guidance, so that the availability and the foreseen system-performance is transparent at any time. For the MT-Engine of the ATLAS –System we decided on a hybrid architecture combining EBMT and SMT at word-based level. For the SMT-component PoS and domain factored models are used, in order to ensure domain adaptability. The document categorization module assigns to each document one or more domains. For each domain the system administrator has the possibility to store information regarding the availability of a correspondent specific training corpus. If no specific trained model for the respective domain exists, the user is provided with a warning, telling that the translation may be inadequate with respect to the lexical coverage. The output of the summarization module is processed in such way that ellipses and anaphora are omitted, and lexical material is adapted to the training corpus. The information extraction module is providing information about metadata of the document including publication age. For documents previous to 1900 we will not provide translation, explaining the user that in absence of a training corpus the translation may be misleading. The domain- and dating restrictions can be changed at any time by the system administrator if an adequate training model is provided.