Machine Translation in ATLAS – Applied Technology for Language Aided CMS - Project

EU CIP-ICT Policy Support Programme
CIP-ICT-PSP-2009-3 Theme 3 Multilingual Web
Pilot Project Type B
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 Zadar, Croatia.
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. i-Publisher will also be used to build two thematic content-driven web sites – i-Librarian and EUDocLib. ATLAS makes use of state-of-the art text technology methods in order to extract information and cluster documents according to a given hierarchy. A text summarization module and a machine translation engine are embedded as well as a cross-lingual semantic search engine. ***The Machine Translation service*** has as main component the statistical paradigm and uses as primary source the Moses System. The main challenge of this service is the domain portability. Statistical systems are highly dependent on training corpora. As large parallel corpora cannot be available for each language pair and any possible domain, which may occur in ATLAS System, we consider following approach:

- We presume that documents are previously categorised to a certain domain
- For each domain we store information regarding domain specific corpora for each involved language pair. According to this information we offer to the user a certain confidence interval fort the translation quality.
- We follow the recent state-of-the art technique of implementing factored models by mixing the training of language models on small domain specific corpora with large general corpora.
- Patterns, characteristic for technical or law texts, are processed through an EBMT-component.

Currently we are developing and improving translation models and we are selecting the optimal parameter setting for each language pair. A first demo will be available in June 2011.

Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.)
Proceedings of the 15th Conference of the European Association for Machine Translation, p. 325
Leuven, Belgium, May 2011