ACCURAT: Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation

Seventh Framework Programme
Call FP7-ICT-2009-4, ICT-2009.2.2: Language-based interaction
Small or medium-scale focused research project (STREP)
Grant Agreement n° 248347
http://www.accurat-project.eu

List of partners
Tilde, Latvia (coordinator)
University of Sheffield, Computer Science Department, NLP Group, UK
University of Leeds, Centre for Translation Studies, U
Institute for Language and Speech Processing, Greece
University of Zagreb, Faculty of Humanities and Social Sciences, Department of Linguistics, Croatia
DFKI, LT Lab, Germany
Romanian Academy, Research Institute for Artificial Intelligence, Romania
Linguatec, Germany
Zemanta, Slovenia

Project duration: : January 2010 – June 2012

Summary

The ACCURAT project is researching methods and techniques to overcome one of the central problems of machine translation (MT) – the lack of linguistic resources for under-resourced languages (Croatian, Estonian, Greek, Latvian, Lithuanian, Romanian and Slovenian) and domains. The scientific objectives of the ACCURAT project are to:

- **create comparability metrics** – to develop a methodology and determine criteria to measure the comparability of source and target language documents in comparable corpora;
- **research methods for the alignment and extraction** of lexical, terminological and other linguistic data from comparable corpora;
- **research methods for automatic acquisition** of a comparable corpus from the Web;
- **measure improvements** from applying acquired data against baseline results from statistic and rule-based MT systems.

The project is in the middle now and several important results are obtained:

- an initial set of criteria of comparability are identified and comparability metrics is implemented to compute multidimensional features that show the level of comparability;
- initial methods have been developed to identify comparable documents on the Web. These methods are used to gather large collections of comparable documents for all project language pairs;
- existing alignment strategies designed for parallel, comparable and non-comparable corpora are evaluated;
- a method for automatic generation of parallel and quasi-parallel data from any degree of comparable corpora ranging from parallel to weakly comparable is developed.

Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.)
Proceedings of the 15th Conference of the European Association for Machine Translation, p. 323
Leuven, Belgium, May 2011