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, UK
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

Lack of sufficient parallel data for many languages and domains is currently one of the major obstacles to further advancement of automated translation. The ACCURAT project addresses this issue by researching methods for using comparable corpora as resources for machine translation (MT). The objectives of the ACCURAT project are to develop methods allowing to measure the comparability of source and target language documents in comparable corpora; to research methods for the alignment and extraction of lexical, terminological, and other linguistic data from comparable corpora; to research methods for automatic acquisition of a comparable corpus from the Web and to analyse how acquired data can improve MT systems. The project particularly targets a number of under-resourced languages, i.e., Croatian, Estonian, Greek, Latvian, Lithuanian, and Romanian, and evaluates applicability of data extracted from comparable corpora for adapting MT to specific narrow domains.

Several novel approaches for building comparable corpora from the Web have been researched and evaluated for under-resourced languages including: (1) monolingual crawling and bilingual pairing of news texts and (2) focused monolingual crawling of narrow domain texts using seed terms and URLs.

ACCURAT has developed comparability metrics which identify similar documents in comparable corpora and indicate their degree of similarity by computing a comparability score. Tests performed on a gold standard show that scores obtained from the metrics reliably reflect comparability levels, as the average scores for higher comparability levels are always significantly larger than for lower levels.

ACCURAT also proposes new methods for extraction of parallel data from comparable corpora. These methods are implemented in an open source ACCURAT Toolkit. The toolkit identifies (maps) and extracts parallel sentences, translation dictionaries, bilingual terminology, and named entities.

Data collected and extracted using ACCURAT tools are being integrated into baseline SMT systems (trained on available parallel data) to evaluate the applicability of ACCURAT tools for improving the quality of MT. Several successful proof-of-concept experiments for narrow domains were carried out showing that even small amounts of parallel domain specific data will help improve a SMT system.