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

ACCURAT consortium

e-mail: project@tilde.lv

www.accurat-project.eu
Challenge of Data Driven MT

- Rapid development of data-driven methods for MT
- Automated acquisition of linguistic knowledge extracted from huge parallel corpora provide an effective solution that minimizes time- and resource-consuming manual work

- Applicability of current data-driven methods directly depends on the availability of very large quantities of parallel corpus data
- Translation quality of current data-driven MT systems is very low for under-resourced languages and domains
Comparable Corpora

- Non-parallel bi- or multi-lingual text resources are much more widely available than parallel translation data.
- Documents are gathered according to a set of criteria (e.g. proportion of texts of the same genre in the same domains in the same period) in two or more languages that contains overlapping information.
- Examples - multilingual news feeds, web pages, Wikipedia articles, blogs etc.
ACCURAT Mission

To significantly **improve MT quality**

- for **under-resourced** languages and narrow domains
- by researching novel approaches how **comparable corpora** can compensate for a shortage of linguistic resources
Use Cases

- Adjusting MT to narrow domain
  *Automotive engineering, assistive technology and data processing domains*

- Application for Web authoring
  *Blog and social networking (Zemanta application)*

- Using SMT in software localization
  *Increasing efficiency in localization, integration with CAT tools*
ACCURAT Focus Areas

ACCURAT will focus on MT areas where scarcity of data poses a major challenge:

- Under-resourced languages
- Narrow domains

AND

ACCURAT methods will be:

- Adjustable to new languages and domains
- Language independent where possible
Language Coverage

- Focus on under-resourced languages: Latvian, Lithuanian, Estonian, Greek, Croatian, Romanian and Slovenian
- Major translation directions like English-Lithuanian, English-Croatian, German-Romanian
- Minor translation directions like Lithuanian-Romanian, Romanian-Greek and Latvian-Lithuanian
Key Innovations

- Novel methods for assessing comparability of corpora
- Novel methods for gathering comparable corpora from the web
- Novel techniques for aligning text segments at various levels across languages using texts from comparable corpora
- Novel ways of exploiting comparable corpora to improve quality of MT
Work Plan

- **WP1**: To create comparability metrics - to develop the methodology and determine criteria to measure the comparability of source and target language documents in comparable corpora (M3-M24)

- **WP2**: To elaborate advanced techniques for extraction of lexical, terminological and other linguistic data from comparable corpora to provide training and customization data for MT (M3-M23)

- **WP3**: To develop, analyze and evaluate methods for automatic acquisition of a comparable corpus from the Web (M1-M22)

- **WP4**: To measure improvements from applying acquired data against results from baseline SMT and RBMT systems (M7-M26)
Work Plan

- WP5: To **evaluate and validate** the ACCURAT project results in three practical applications (M7-M30)
- WP6: To **disseminate** project results and to transfer the project knowledge, technologies, lessons learned and best practices to interested communities and thus to ensure their worldwide impact and long-term sustainability (M1-M30)
- WP7: To **coordinate** the project and provide administrative and financial **management** (M1-M30)
Iterative Approach

- 4 evaluation points: M10, M16, M23, M26
Milestones

- **Initial comparable corpora** (M3)
  - Tools for collecting comparable corpora from the Web (M22)
  - Multilingual comparable corpora (M22)

- **Initial comparability metrics** (M6)
  - Criteria and metrics of comparability and parallelism (M24)

- **Application of existing alignment methods** (M6)
  - Alignment and extraction methods for comparable corpora (M20)

- **Baseline SMT systems** (M9)
  - Improved MT systems (M26)
  - Adjusted MT systems in applications (M30)
Key Results

- **Comparability metrics** developed and tools provided
- **Comparable corpora** for under-resourced languages collected and tools provided
- Methods and tools for **multi-level alignment** from comparable corpora developed
- Methods for using comparable corpora in both **SMT** and **RBMT** developed
- Proven **application scenarios** prepared

Strong increase in MT quality for under-resourced languages and narrow domains
Partners

Tilde (Coordinator)  Latvia
University of Sheffield  UK
University of Leeds  UK
Athena Research and Innovation Center in Information Communication and Knowledge Technologies (ILSP)  Greece
University of Zagreb, Faculty of Humanities and Social Sciences  Croatia
DFKI  Germany
Institute of Artificial Intelligence  Romania
Linguatec  Germany
Zemanta  Slovenia
Complementing Competencies

- Industry experience
- Language expertise
- Research excellence
ACCURAT project has received funding from the EU 7th Framework Programme for Research and Technological Development under Grant Agreement N° 248347

Project duration: January 2010 - June 2012

Contact information:
Andrejs Vasiljevs
andrejs@tilde.lv
Tilde, Vienibas gatve 75a, Riga
LV1004, Latvia

www.accurat-project.eu