ACCURAT - Analysis and evaluation of comparable corpora for under resourced areas of machine translation

Andrejs Vasiļjevs

Tilde

Language Technology Days, Luxembourg

March 22-23, 2010

Do you have

translation?

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

Challenge of Data Driven MT

- 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

- The non-parallel bi- or multi- lingual text resources are much more widely available than parallel translation data
- Collection of documents that 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
  ACCURAT application for blog and social networking
- 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

BUT

ACCURAT methods will be adjustable to the new languages and domains and 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

State of the art

Comparability metrics

- No agreement on the **degree of similarity** that documents in comparable corpora should have
- No agreement on the **criteria for measuring parallelism** and comparability
- Few studies to assess similarity of content using frequency lists, named entity recognition, term extraction comparison

Multi-level alignment methods and information extraction

- Initial research for applying CLIR techniques in **selection process** for widely used languages
- **Alignment methods** designed for parallel texts have poor results on comparable corpora
- Initial research on application of word-overlap filter together with a constraint on the ratio of lengths and Maximum Entropy classifier **for sentence alignment**

State of the art

Building comparable corpus from the Web

- **General methods** and techniques for building corpora from the and extracting **parallel corpora** for well-resourced languages
- Munteanu and Marcu (2005) have built techniques to **extract parallel sentences from** comparable corpora from multilingual **news** sites
- Tools to collect comparable corpora from the Web **for under-resourced languages do not exist**

Comparable corpora in MT systems

- Current SMT systems are mostly based on translation models derived from **parallel corpus data**
- Few initial studies **for few languages**
- The latest research has shown that adding data from comparable corpora **improves the system's performance** in view of un-translated word coverage

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

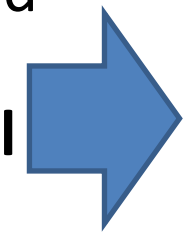
- WP1: To create **comparability metrics** – methodology and criteria to measure the comparability of source and target language documents in comparable corpora
- WP2: To elaborate advanced techniques for **extraction** of lexical, terminological and other linguistic data from comparable corpora to provide training and customization data for MT
- WP3: To develop, analyze and evaluate methods for **automatic acquisition** of a comparable corpus from the Web
- WP4: To **measure improvements** from applying acquired data against results from baseline SMT and RBMT systems

Work Plan

- WP5: To **evaluate and validate** the ACCURAT project results in **three practical applications**
- 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
- WP7: To **coordinate** the project and provide administrative and financial **management**

Key Results

- **Comparability metrics** proposed 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

SEVENTH FRAMEWORK PROGRAMME

First results: initial comparable corpora

Domain	Genre	Coverage
International news	Newswires	20%
Sports	Newswires	10%
Admin	Legal	10%
Travel	Advice	10%
Software	Wikipedia	15%
Software	User manuals	15%
Medicine	For doctors	10%
Medicine	For patients	10%

- parallel texts
- strongly comparable texts
- weakly comparable texts

Partners

Tilde	Latvia
University of Sheffield	UK
University of Leeds	UK
ILSP	Greece
University of Zagreb	Croatia
DFKI	Germany
Institute of Artificial Intelligence	Romania
Linguatec	Germany
Zemanta	Slovenia

Duration

January 2010 – June 2012

Complementing Competencies

Advisory User Group

- Represents both target groups of ACCURAT:
  - the research community
  - potential users
- The ACCURAT Advisory User Group:
  - Providing feedback from research and user communities
  - Participation in evaluation
  - Synergy with other projects and activities
  - Dissemination channel
  - Ensuring awareness about the project results

More information

- www.accurat-project.eu
- ACCURAT stand at the EC Village at LREC 2010
- Workshop on Methods for the automatic acquisition of Language Resources and their evaluation methods
  (LREC 2010, May 23)
  Organized by FLaReNet, ACCURAT, PANACEA and TTC
- 3rd Workshop on Building and Using Comparable Corpora (LREC 2010, May 22)
- 4th International Conference *Human Language Technologies – the Baltic Perspective* (October 7-8, 2010)
  Organized by Tilde, LU MII, ACCURAT, LetsMT!

Beyond ACCURAT

- Using the whole Web as a dynamic, multilingual comparable corpus
- Self-improving MT systems by dynamically collecting and processing Web data
- Self-adaptive MT for various domains and application needs
- Extracting semantic information from corpora: ontologies and knowledge bases

Thank you!

www.accurat-project.eu

andrejs@tilde.lv