Platform for Online Sharing of Training Data and Building User Tailored MT

Project ID card

- Funded under: The Information and Communication Technologies Policy Support Programme
- Area: CIP-ICT-PSP.2009.5.1 Multilingual Web: Machine translation for the multilingual web
- Total cost: €3.34m
- EU contribution: €1.67m
- Project reference: 250456
- Execution: From 01/03/2010 to 31/08/2012
- Project status: Running
- Contract type: The Information and Communication Technologies Policy Support Programme PB Pilot Type B

Challenge:

In recent years, statistical machine translation (SMT) has become the leading paradigm for machine translation. The quality of SMT systems largely depends on the size of training data. Since the majority of parallel data is in major languages, SMT systems for larger languages are of much better quality compared to systems for smaller languages. This quality gap is further deepened due to the complex linguistic structure of many smaller languages. Languages like Latvian, Lithuanian and Croatian (to name just a few) have complex morphological structure and free word order. To learn this complexity from corpus data, much larger volumes of training data are needed. Current systems are built on the data accessible on the web, but it is just a fraction of all parallel texts. Most of them still reside in the local systems of different corporations, public and private institutions, and desktops of individual users.

Proposed solution

To fully exploit the huge potential of existing open SMT technologies we propose to build an innovative online collaborative platform for data sharing and MT building. This platform will support upload of public as well as proprietary MT training data and building of multiple MT systems, public or proprietary, by combining and prioritizing this data. LetsMT! will extend the use of existing state-of-the-art SMT methods that will be applied to data supplied by users to increase quality, scope and language coverage of machine translation.

The LetsMT! platform will be of particular significance for users of smaller languages where the currently available MT solutions are of unsatisfactory quality, due to the limited amount of parallel texts available for training, or where they do not exist at all. This platform will enable user collaboration in identifying and sharing of parallel texts that currently are not publicly available.

Target groups:

LetsMT! project will provide MT solutions for European citizens and businesses allowing more efficient usage of multilingual content. Specifically, LetsMT! platform will target the localisation and translation industry content (language services providers, enterprises and organisations with multilingual content, freelance translators, etc), and the audience of business and financial news; at the same time, it will be of interest for a variety of users: web users in general, speakers of less-covered languages, academic community and others.

Benefits:

For actors in the localisation and translation industry, LetsMT! will provide facilities for training of SMT systems on their data and generating custom SMT solutions to be used by localisation service providers as well as their clients.

For users of business and financial news, LetsMT! will provide free and instant MT services, with emphasis on less-covered languages.

For all holders of linguistic content, the LetsMT! platform will easily build MT services using their specific content.

The result:

LetsMT! will provide a platform that supports the following features:

- Uploading of parallel texts for users that will contribute their content
- Directory of web and offline resources gathered by LetsMT! as well as links provided by users to other sources not included in the LetsMT! repository
- Automated training of SMT systems from specified collections of training data
- Custom building of MT engines from selected pool of training data, for larger donors or paying customers
- Custom building of MT engines from proprietary non-public data, for paying customers
- MT evaluation facilities

The solution will deliver the following core functionality:

- **website for upload** of parallel corpora and building of specific MT solutions
- **website for translation** where source text can be typed and translated
- **translation widget** provided for inclusion into websites to translate their content
- **browser plug-ins** that will provide the quickest access to translation
- **integration in CAT tools** and other applications

Impact:

LetsMT! results will have significant impact in achieving objectives of CIP-ICT-PSP Work Programme, namely:

- Significant increase in available language resources for training of SMT systems
- Improved quality of SMT, especially for smaller languages
- Increase in language coverage for MT
- Diversification of free MT by tailoring for specific domains or user requirements
- Significant increase in usage of MT on the Web and in applications through LetsMT! translation widgets, plug-ins and MT web-service
- Much wider use and greater impact of available open-source SMT technologies
- Collaborative involvement of different stakeholders from the public sector, SMEs, universities, research and education community

Contact information

BERZINS Aivars
VIENIBAS GATVE 75 A, 1004 RIGA
REPUBLIC OF LATVIA
Tel. +37167605001
Fax. +37167605750
E-mail: aivars.berzins@tilde.lv