LetsMT!

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

www.letsmt.eu/

Project no. 250456

Annual Public Report

Version No. 1.0

15/11/2010

Contents

1. PROJECT CHALLANGE

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 a 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 this is just a fraction of all parallel texts. Most parallel texts still reside in the local systems of different corporations, public and private institutions, and on the desktops of individual users.

2. PROJECT OBJECTIVES

To fully exploit the huge potential of existing open SMT technologies, the main objective of the LetsMT! project is to build an innovative online collaborative platform for data sharing and machine translation (MT) building. LetsMT! will be a collaborative platform that thrives on resources contributed by its users. It will have a major breakthrough effect regarding the availability of parallel language resources and, consequently, MT services of good and acceptable quality for less-covered languages where the current MT systems perform poorly due to limited availability of training data.

This platform will support the upload of public as well as proprietary MT training data and the 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. LetsMT! will provide a platform that supports the following features:

- Uploading of parallel texts for users that will contribute their own content
- Directory of web and offline resources gathered by LetsMT! as well as links provided by users to other content sources not included in the LetsMT! repository
- Automated training of SMT systems from specified collections of training data
- Custom building of MT engines from a 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 uploading** parallel corpora and building 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

3. SUMMARY OF ACTIVITIES

During the first months of the project, foundations for development work have been put in place, cooperation between the project partners has been established and the first results of the work of the consortium have been achieved. Management procedures were established and a project management plan created. Project partners established a common understanding and approach for the project implementation.

Project LetsMT! is divided between 6 major activities with supplementing subtasks. Each activity is dedicated to a specific element of the LetsMT! platform such as infrastructure building, data collecting, training etc.

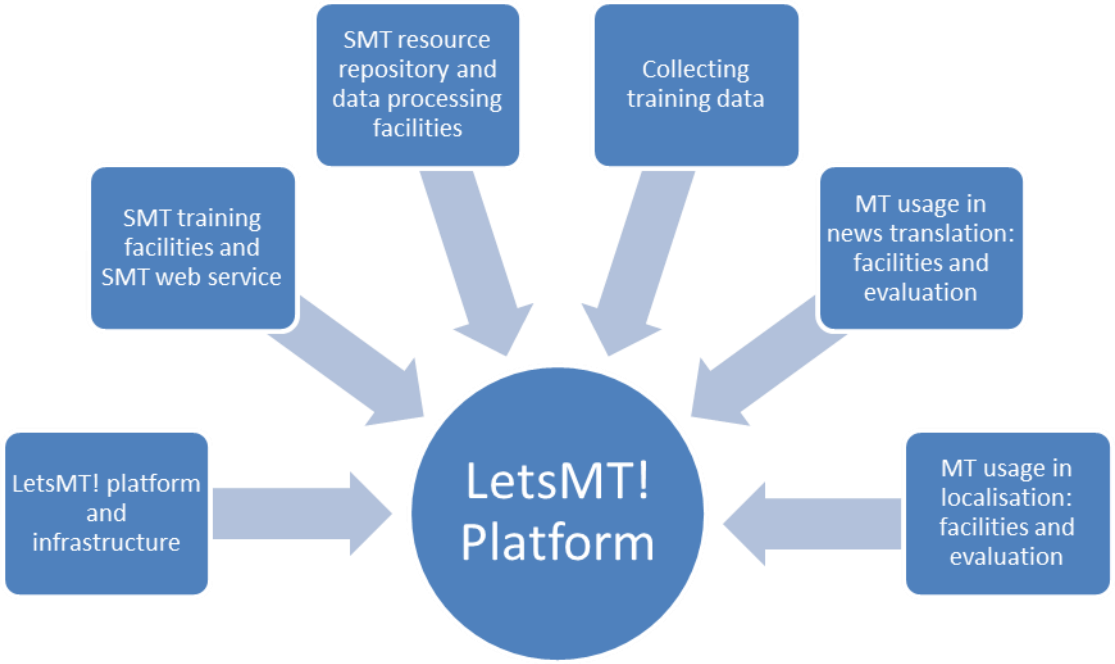

Figure 1. LetsMT! structure

One of the first tasks in the project was a requirements analysis. At this phase we have conducted interviews with potential users of the LetsMT! platform to indicate major factors and user requirements. The gathered requirements have been analyzed, prioritized, and a requirements analysis was written. The evidence that is used in this requirements analysis is collected from two sources. An internal source in terms of two project partners' detailed description of use scenarios and an external view on user requirements provided by collecting information from a wide range of end users.

The next step to establish the foundation for the LetsMT! platform was to analyse data formats. One of the key functions of the LetsMT! platform is to provide the possibility to train domain specific SMT models tailored towards specific needs of its users. For this appropriate data resources are required. The scope of the data format task is to develop facilities to store, process and manage resources coming from LetsMT! users. Based on users' requirements and

the nature of the LetsMT! platform, we have created a report on the specification of data formats supported by the platform. The purpose of this report was to specify the data formats that will be used internally when storing parallel and monolingual corpus data in the repository and to list data formats that will be supported for the user contributions.

The functional specification has been created based on the analysis of requirements. It describes the features of the LetsMT! system and the high-level system design necessary to support further development and implementation of the platform. Development work on the implementation of the LetsMT! platform is started according to the functional specification document. It is planned to release the first prototype of the LetsMT! platform in early 2011.

One of the core tasks of the project is to collect training data. Volume of training resources has a major influence on the quality of MT. During the first 8 month of the project we have started work on resource collection tasks. Primarily parallel texts are needed as training data, but large monolingual corpora can also be used to produce language models for target languages. In order for texts to be useful as training data it is important to have access to all available information about their origin, subject domain (domain classification) and text formats.

Administration of training data is an important subject as the MT quality will depend on high quality training data and on the ability to select the most appropriate data for the training process.

The initial observations show that much data is already available but besides the collection of domain specific data we might be left with the task of obtaining and populating the system with more monolingual and parallel data for general language for some of the languages in focus.

4. TARGET USERS

During interviews, conducted in order to gather user requirements, several types of users were identified with specific requirements for the LetsMT! platform. A summary of the user types within the different user groups and their expected interaction with the LetsMT! system is outlined in the table below.

User type	Description
Individual Internet users	
Anonymous Internet User	A person browsing the Web; who wants to test SMT or hopes to receive better machine translation than provided by other publicly available MT systems like Google Translate. This user does not want to make an effort, just use the service to either translate a desired text or to compare with other SMT engines. This user is not expected to contribute corpora
Business news reader	Wants to read international financial news in “smaller” native EU languages. Foreign language skill is limited; therefore LetsMT! is used to gist English news in local language and vice versa. Requirements for translation quality are not very high, however it is expected that terminology and the essence of news will be translated correctly and

User type	Description
	understandably. This user is not expected to contribute corpora.
MT enthusiast	A person who is aware of various available MT tools and technologies. Attraction to LetsMT! is based on the possibility to directly influence the SMT engine and review the SMT system training and translation logs. This MT enthusiast is ready to contribute corpora in order to achieve better SMT results. Receiving praise for contributing or belonging to some elite group would be considered a bonus. Might submit poor quality comparable corpora.
Localization and translation industry company	
Translator	Needs to use SMT because of organization's workflow. May be very skeptical of MT results. Resistant to change. Wants to use SMT seamlessly integrated in daily routines. CAT tool integration is preferred. Very simple and quick on-line tool could be acceptable that supports file formats used in translation projects.
Translation Project Manager	Wants to reduce translation project cost by reducing time translators spend on initial translation. Would be ready to contribute high-quality corpora/translation memories in order to improve SMT system quality. Needs to have good control and quality measurements of trained SMT systems. Different clients have different domains, vocabularies, tools, translation styles. Thus a single SMT system does not provide the necessary quality of translation and post-editing of machine translated text takes the same or even longer time. Will need lots of specific SMT systems for each project type or customer. Would use only a few SMT systems simultaneously, but needs lots of historical data to build SMT systems on-demand whenever specific project starts.
Localization Company Manager	Have high quality corpora which are most likely protected by IPR or confidentiality agreements. Highly aware of IPR and the need to protect organization-specific knowledge (competitive advantage). SMT usage in current translation workflow and tools must be cost and time efficient. Currently maintenance and even evaluation of SMT feasibility is very expensive due to lack of knowledge of technologies involved and infrastructure requirements.
Web developers	
Web developer	Web developer who needs to create multi-lingual sites, but does not have necessary resources to provide high quality translation.

User type	Description
	Commonly will use LetsMT! for prototyping of web sites in different languages. If the quality of the translation is accepted by the client in general, the web developer could provide minor fixes and improvements of translation.
University education and research community	
Researcher	Member of educational and research organization or translation and localization industry organization investigating options available in SMT field. Most interested in quality of translation and possibilities to control SMT system training. Has some parallel and monolingual corpora available and is interested in improvements in translation quality after corpora submission. This user will want to tweak every possible SMT system option and compare results of these tweaks. Could be a decision maker or contribute to decision making about use of LetsMT! services.
Research organization leader	Owns large amounts of mono and bi-lingual corpora. Is interested in theoretical aspects of results of SMT systems and how different input data influence quality of translations. Research organizations most likely will have many different SMT systems and will re-train SMT systems frequently. Would benefit from using LetsMT! platform as no investment in infrastructure is required. Funding possibilities are limited.
Translation automation solution developers and providers	
Translation Solution Product manager	As user community constantly requires MT solution integration solution developer would benefit from integration of LetsMT! platform into their products/solutions. Also many competitors have introduced MT modules in their solutions. Installation and maintenance of MT requires infrastructure and skilled specialists in MT are scarce. Even access to public SMT systems through products could be seen as beneficial to customers as trial of MT integration. More user-tailored solutions could be sold as a separate service.

5. USAGE SCENARIOS

There are a number of freely available translation systems on the Internet; however none of those allow users the options to choose between domain–specific systems. Users cannot directly influence SMT systems by uploading corpora in user's possession for achieving better quality of translations for their needs. LetsMT! will address these shortcomings.

Users will be able to upload corpora into the system, using a widely accessible client application, for example, by using a web page interface. Users will be provided with an effective way of searching, navigating and selecting a trained SMT system to use. And, of course, translate texts using one of available LetsMT! trained SMT systems.

In particular two specialized usage scenarios will be supported by the LetsMT! system: translation project in localization industry companies (see Figure 2. Localization project process overview below) and financial news machine-translation.

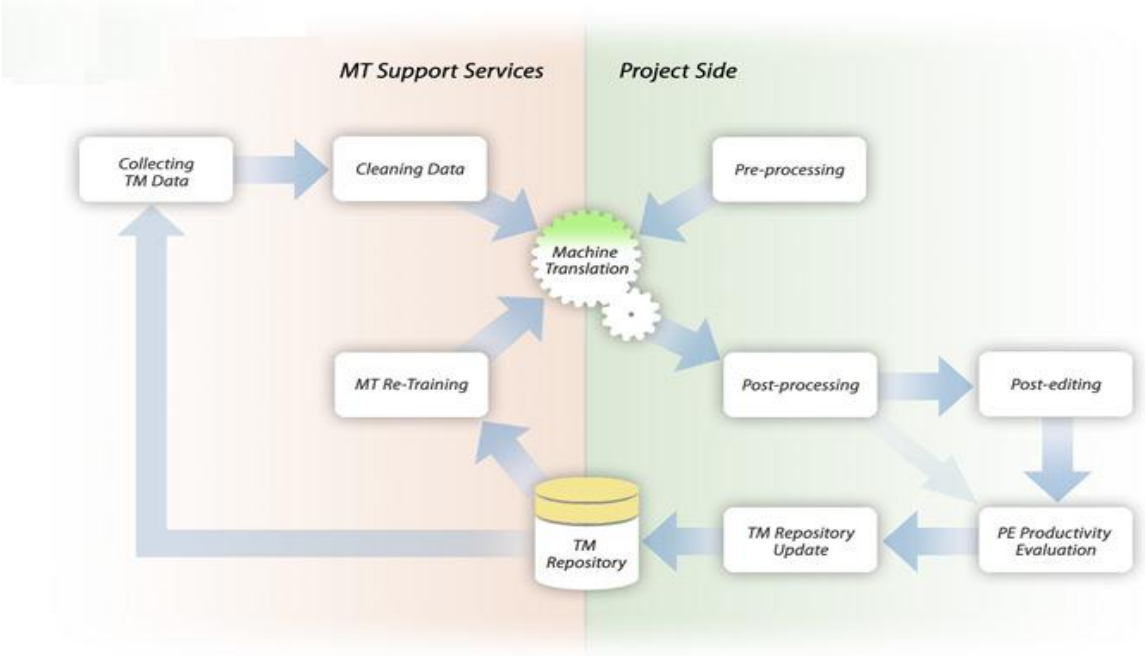

Figure 2. Localization project process overview in localization industry company

6. DISSEMINATION

Dissemination Strategy

Dissemination Plan and Visual identity

The dissemination plan was completed at M3 defining all necessary means of dissemination of information about the LetsMT!: target groups, dissemination channels, general visual identity, public web-site design, posters, flyers, t-shirts, participation on key conferences, LetsMT! events, public showcases, web presence, appearance in scientific journals and conferences, issuing announcements, social-networks presence etc.

The visibility of the LetsMT! project is assured by a unique visual identity (logo) that helps in recognizing the project among similar projects. The visual identity was designed and applied to all possible channels of dissemination such as the public web site, presentation template, leaflets, posters, but also t-shirts and more unconventional channels such as e.g. video lectures, Wikipedia articles or social networks.

The LetsMT! website (www.letsmt.eu/) is one of the main project communication tools. Alternative addresses are www.letsmt.org, www.letsmt.com.

Dissemination to the scientific community

Dissemination to the scientific community is based on a bilateral exchange of information by consortium partners with major scientific institutions as well as communication of project achievements in conferences and through publications.

Dissemination of LetsMT! results in the scientific community is carried out through presentations of research methodologies, strategies, and outcomes in industry conferences.

Project news and findings are being reported and posted primarily to the project web site, but a noticeable presence is also being planned on international online science and technology portals. Ongoing dissemination to the wider academic community has already taken place in peer reviewed international publications where papers with results from LetsMT! project have been accepted for publishing.

Initial flyers and posters have been produced and exhibited on several occasions at various conferences and events. Presentations on LetsMT! were held in conjunction with the LREC2010 conference (Valletta, Malta), the NooJ2010 conference (Komotini, Greece), Translingual Europe 2010 (Berlin, Germany), the FASSBL2010 conference (Dubrovnik, Croatia), and the Baltic HLT conference (Riga, Latvia). LetsMT! was one of the organizers of the Baltic HLT conference.

Dissemination to the industry

It is foreseen to promote the project's innovative technologies in the framework of national and international conferences, exhibitions and events most attractive to the language industry (such as localisation and/or translation industry events etc.).

A three day presentation at the EC Project Village was held in conjunction with the major scientific but also language technology industry conference in the field (LREC2010, Valletta, Malta) with an exhibition booth and the dissemination of print materials.

The LetsMT! Project was presented by A. Vasiljevs on a session about the European industry and academic cooperation for innovation and leadership in language technologies at the Baltic IT&T2010 conference in Riga, Latvia (April 22, 2010).

LetsMT! was positioned as a cloud-based service for MT generation at Multilingual Europe 2010, International Conference on Advance Translation Technology for Multilingual Europe, Berlin, Germany, 2010-07-15 (see video lecture at http://videolectures.net/translingeu2010_vasiljevs_tcb/)

Publications of the project team in the period February to November 2010

Papers

Vasiļjevs, A., Gornostay, T. and Skadiņš, R. LetsMT! Platform for Online Sharing of Training Data and Building User Tailored Machine Translation, Human Language Technologies – The Baltic Perspective, Proceedings of the Fourth International Conference Baltic HLT 2010, Frontiers in Artificial Intelligence and Applications, vol. 219, Riga, Latvia, October 7–8, 2010 (doi: DOI: 10.3233/978-1-60750-641-6-133)

Conference presentations

Vasiljevs, A. LetsMT! – Towards Cloud-Based Service for MT Generation, Translingual Europe 2010, Berlin, Germany, July 15, 2010. (see video lecture at http://videolectures.net/translingeu2010_vasiljevs_tcb/)

Vasiļjevs, A., Gornostay, T. and Skadiņš, R. LetsMT! Platform for Online Sharing of Training Data and Building User Tailored Machine Translation, The Fourth International Conference Human Language Technologies — the Baltic Perspective, Riga, Latvia, October 7–8, 2010.

Presentation "BIG solutions for small languages" by Indra Sāmīte, TAUS Data Association conference taking place in, Portland, USA October 3-6, 2010.

Jörg Tiedemann, Per Weijnitz Let's MT! — A Platform for Sharing SMT Training Data, Swedish Language Technology Conference (SLTC-2010) in Linköping in October 29, 2010.

7. LetsMT! CONSORTIUM AND CONTACT PERSONS

Tilde SIA
Vienibas gatve 75a
Riga, LV1004
Latvia
Project Coordinator:
Andrejs Vasiļjevs
e-mail: andrejs@tilde.lv
URL: http://www.tilde.eu

UNIVERSITY OF EDINBURGH
Old College,
South Bridge Edinburgh EH8 9YL
UK
Contact person:
Philipp KOEHN
e-mail: philipp.koehn@ed.ac.uk
URL: http://www.statmt.org/ued/

UNIVERSITY OF ZAGREB
Trg maršala Tita 14
HR-10002 ZAGREB,
Croatia
Contact person:
Prof. Marko TADIĆ
e-mail: marko.tadic@ffzg.hr
URL: http://hnk.ffzg.hr/default_en.htm

KØBENHAVNS UNIVERSITET
Njalsgade 80, DK-2300
Copenhagen S
Contact person:
Lene Offersgaard
e-mail: leneo@hum.ku.dk
URL: http://www.humanities.ku.dk/

UPPSALA UNIVERSITY
P.O. Box 635, S-751 26 Uppsala,
Sweden
Contact person:
Dr. Jörg Tiedemann
e-mail: jorg.tiedemann@lingfil.uu.se
URL: http://www.lingfil.uu.se/lingfil_eng/

ZOOROBOTICS – (TRADING NAME SEMLAB)
Zuidpoolsingel 14a
2408ZE Alphen a/d Rijn
The Netherlands
Contact person:
Bram Stalknecht
e-mail: stalknecht@semlab.nl
URL: http://www.semlab.nl/pages/index.jsp?id=170

MORAVIA IT A.S.
Hilleho 4
602 00 Brno
Czech Republic
Contact person:
David Filip'
e-mail: DavidF@MoraviaWorldWide.com
URL: http://www.moraviaworldwide.com/