LetsMT! – Towards cloud-based service for MT generation

Andrejs Vasiljevs

andrejs@tilde.com

Tilde

Translingual Europe 2010, Berlin, 07.06.2010

Data challenge

- **Statistical methods** provide breakthrough in cost-effective MT development

- Quality of SMT systems largely **depends on the size** of training data

- To overcome gap in SMT language and domain coverage and to improve quality much larger volume of training **data is needed**

- Parallel data accessible on the web is **just a fraction** of all translated texts. Most of them still reside in the local systems of different corporations, public and private institutions, desktops of individual users.

Let's MT!

Customization challenge

- Current mass-market and online MT systems are of **general nature** and perform poorly for domain and user specific texts.

- System adaptation is prohibitively **expensive service** not affordable to smaller companies or the majority of public institutions.

- Particulary **localization industry** is not able to fully exploit the data they have.

Let's MT!

Platform challenge

- ❑ Great open source platforms like Moses and GIZA++ make it relatively easy to build MT engine.

- ❑ Still expertise and local infrastructure is needed that is not available for majority of users.

Let's MT!

LetsMT! Vision
Let's advance MT together!

- ❑ To fully exploit the huge potential of existing open SMT technologies to create an innovative online collaborative platform for **data sharing and MT building**.

- ❑ This will be a platform that gathers public and user-provided MT training data and generates multiple MT systems 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.
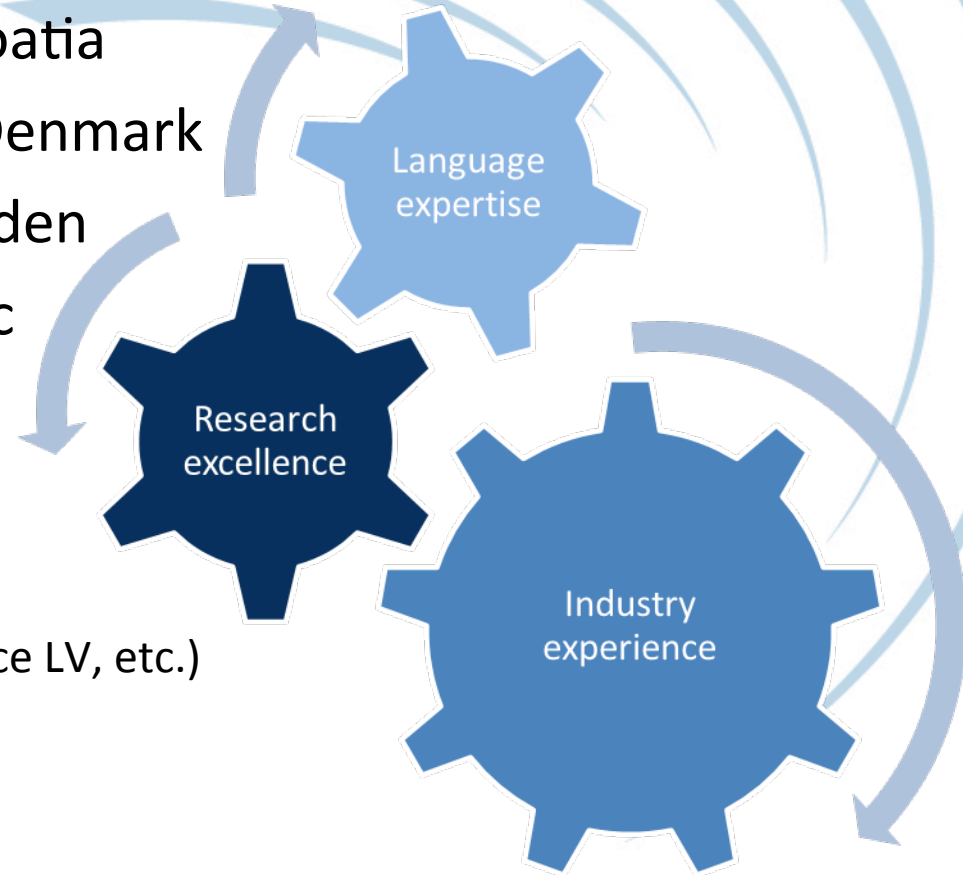
Let's MT!

LetsMT! Vision

- Sustainable user-driven MT factory on the *cloud* providing services for user data sharing, MT generation, customization and running.

Let's MT!

LetsMT! Project ID

- ❏ Funded under: EU Information and Communication Technologies Policy Support Programme

- ❏ Area: CIP-ICT-PSP.2009.5.1 Multilingual Web: Machine translation for the multilingual web

- ❏ Project reference: 250456

- ❏ Execution: From 01/03/2010 to 31/08/2012

- ❏ Project coordinator: Tilde

Partnership with Complementing Competencies

- ❑ Tilde (Project Coordinator) - Latvia
- ❑ University of Edinburgh - UK
- ❑ University of Zagreb - Croatia
- ❑ Kopehagen University - Denmark
- ❑ Uppsala University - Sweden
- ❑ Moravia – Czech Republic
- ❑ SemLab – Netherlands

- \+ Support Group
 (TAUS DA, SDI Media, Patent Office LV, etc.)

Language expertise

Research excellence

Industry experience

Let's MT!

LetsMT! Main Features

- Users will contribute with **user-provided content** by uploading their parallel texts
- **Directory** of web and offline resources gathered by LetsMT! as well as user provided links to other sources that are not yet included in LetsMT! repository
- **Automated training** of SMT systems from specified collections of training data
- Larger donors or customers will be able to specify particular training data collections and build **customised MT engines** from these collections
- Customers will be able to use LetsMT! platform for tailoring MT system to their needs from their **non-public data**
- Users will be involved in **MT evaluation**

Let's MT!

Software Architecture
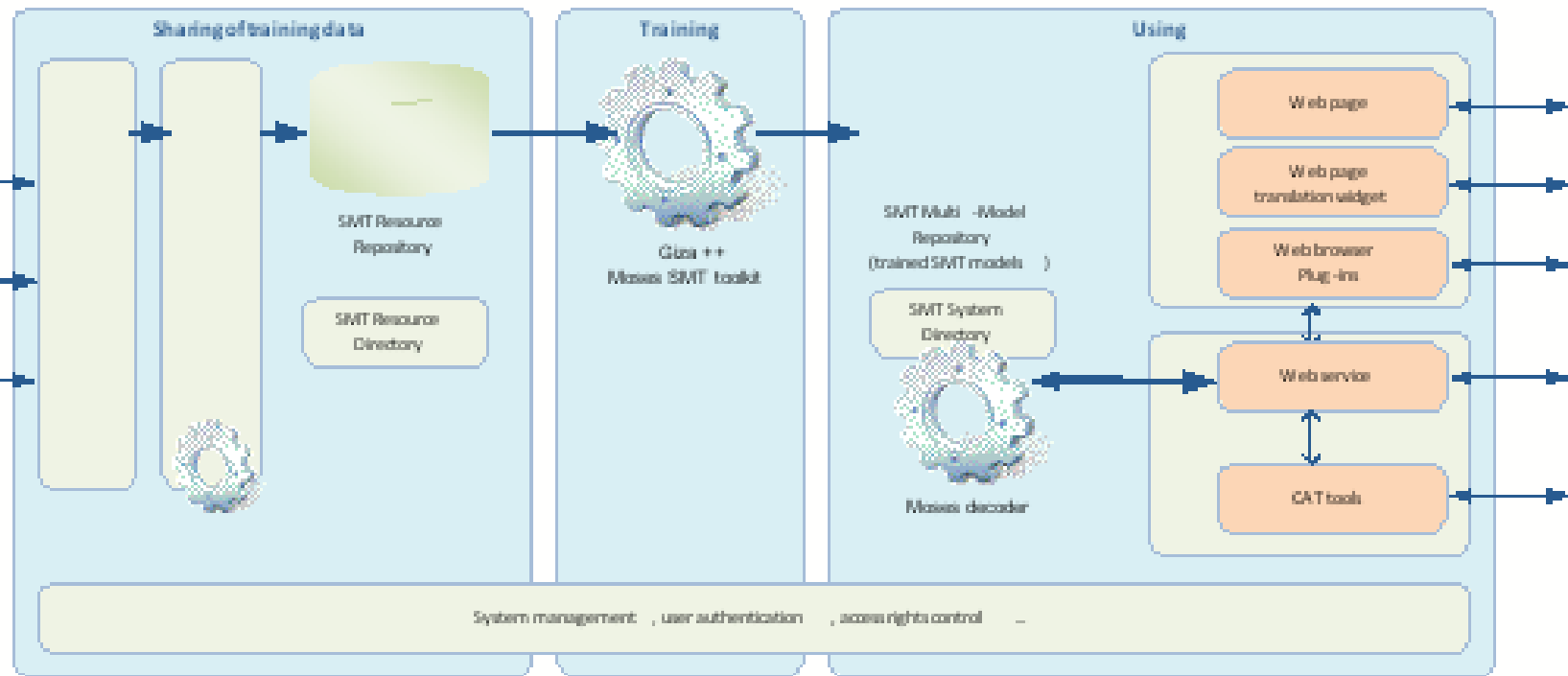
Key Outcomes

- **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 free inclusion into websites to translate their content
- **browser plug-ins or add-ons** that would allow the quickest access to translation
- web service for **integration in CAT tools** and other applications

Let's MT!

Lets MT! main target groups

- ❑ Translation industry
- ❑ Freelance translators
- ❑ Software developers and providers
- ❑ Web developers
- ❑ Public institutions
- ❑ Research community
- ❑ University education
- ❑ General users

Let's MT!

Application Scenarious

- ❑ Online MT service for the **localization and translation** industry
- ❑ Online MT service for global **business and financial news**

- ＋ Showcase for patent translations for gisting purposes

Let's MT!

Key Impact Areas

- Significant increase in available **language resources** for training of SMT systems
- **Improved quality** of SMT, especially for smaller languages
- Increase in **language coverage** for machine translation
- **Diversification** of free MT by tailoring for specific domains or user requirements
- Significant **increase in usage** of MT in web and 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 public sector, SMEs, universities, research and education community

Let's MT!

Thank you and Let's MT!

letsmt.eu