BRIDGING TECHNOLOGICAL GAP
BETWEEN SMALLER AND LARGER LANGUAGES

Andrejs Vasiljevs
Tilde
Pisa Workshop on Multilingual Web
05.04.2011
LANGUAGE DIVERSITY SHOULD BE NURTURED AND TOOLS PROVIDED TO BRIDGE LANGUAGE BARRIERS
UNESCO ON LANGUAGE DIVERSITY IN CYBERSPACE

- Information should be made available, accessible and affordable across all linguistic groups including people who speak minority languages. ICTs shall serve to reduce digital divide and deploy technology and applications to ensure inclusion.

- Creation, preservation and processing of, and access to content in digital form should ensure that all cultures can express themselves and have access to Internet in all languages, including indigenous and minority languages.

//Code of Ethics for the Information Society (Draft)
Survival of smaller languages depends on the outcome of the race between development of Machine Translation and proliferation of larger languages.
ABOUT TILDE

► Tilde – Language technology and localization company
► Offices in Riga (Latvia), Vilnius (Lithuania), Tallinn (Estonia)
► 115 employees, including 3 PhDs and 6 PhD candidates/students in Research department
► Expertise in translation technologies, terminology management and in languages of the Baltic countries
MACHINE TRANSLATION AT TILDE

- Rule based MT in development since 1998
- Very time and resource consuming manual work of software experts and linguists
- No national or EU funding was available
- Tilde’s English-Latvian and Latvian-Russian RBMT released in 2007
- First on the market but reasonable quality only for simpler texts
- Switching to data-driven statistical methods in 2008
- Heavy participation in EU R&D to foster MT development
This is the second of four workshops that survey and share information about currently available best practices and standards that can help content creators and localizers address the needs of the multilingual Web, including the Semantic Web. They also provide an important opportunity to identify gaps that need to be addressed. The workshop is also designed as an opportunity for participants to network and share information between and across the various different communities involved in enabling the multilingual Web.

About Tilde Translator

Tilde Translator is a machine translation tool developed by Tilda. The Beta version provides translation from English to Latvian and from Latvian to English. We are working on providing translations in other language combinations too.

Please note that even the best machine translation cannot be compared to text that has been translated by a human being. Often the content and grammar of texts translated by machine is inaccurate and can contain errors. Yet we hope that Tilde Translator will be useful for you and that it will be a good way to grasp the idea of texts and to break down language barriers.

Machine Translation Blog

Tilde Translator Presented at Research Workshop: Machine Translation and Morphologically-rich Languages. In the end of
Partneribās ir tiešķīga pakalpojuma Microsoft Translator galvenais aspekts. Microsoft Translator mašintulkojuma pamattehnoloģija tika veidota, pamatojoties uz vairāk nekā desmit gadu gaitā veiktiem pētījumiem Microsoft Research centrā, un mēs uzskatām, ka, lai nodrošinātu pasaules klasē tulkotās ārējās pieminekļu kvalitāti un tulkotās materiāla pareizu atspoguļojumu.

"Strādājam pie pakalpojuma Microsoft Translator latviešu-angļu un angļu-latviešu programmas uzlabojumiem. Microsoft Research sākotnēji ir uzņēmumu Tilde. Tilde sniedza ideāliskus par vairākiem tehnoloģijām, kas bija aizstājams ar mašintulkojumu, dažām un spēcīgajām latviešu valodai raksturīgajiem rīkiem un tehnoloģijām. Tilde ir izveidojusi uzbūvi latviešu valodas tulkojumu pēc kalpojuma Microsoft Translator."

Tilde

Sasniegs Tilde tika dibināta 1991. gadā, un tas ir valodas Baltijas IT uzņēmums ar speciālizētu lokalizāciju, vairākvalodu un interneta programmatūru, kā arī inovatīvu līdzekļu strādā Baltijas valstu valodā ieguves tehnoloģijā. Tilde ir viens no Baltijas valstu valodā izveidojusi tehnoloģiju, kas ir ļoti izdevīgs pasākums izpildītāju valodā, atvainojoties apvietījumu un privātām uzņēmumā ar braucējiem Rīgā, Tallinā un Vīržā. 
CHALLENGE OF DATA DRIVEN MT

- Rapid development of data driven methods for MT
- Automated acquisition of linguistic knowledge extracted from parallel corpora replace 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 low for under-resourced languages and domains
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.
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.

- Particularly **localization industry** is not able to fully exploit the data they have.
PLATFORM CHALLENGE

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

► Still expertise and local infrastructure is needed that is not available for majority of users.
SOME STRATEGIES TO BRIDGE THE GAP

► Encourage users to share their data
► Involve users in MT improvements
► Use other kind of multilingual data beyond parallel texts
To better exploit the huge potential of existing open SMT technologies to create an innovative online collaborative platform for data sharing and MT building.

LetsMT! is building a platform that gathers public and user-provided MT training data and generates multiple MT systems by combining and prioritizing this data.

LetsMT! extends the use of state-of-the-art SMT methods to data supplied by users increasing quality, scope and language coverage of machine translation.
Sustainable user-driven MT factory on the cloud providing services for user data sharing, MT generation, customization and running.

Let'sMT! Project
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

- Tilde (Project Coordinator) - Latvia
- University of Edinburgh - UK
- University of Zagreb - Croatia
- Copenhagen University - Denmark
- Uppsala University - Sweden
- Moravia – Czech Republic
- SemLab – Netherlands
USER SURVEY: IPR OF TEXT RESOURCES IN INTERVIEWEE ORGANIZATIONS

- 37% no reply
- 23% interviewee has IPR
- 22% interviewee has restricted/partial IPR
- 18% interviewee has no IPR
USER SURVEY: WILLINGNESS TO SHARE DATA

- 40% no reply/interviewee has no data now
- 23% yes
- 16% perhaps
- 21% no
SOFTWARE ARCHITECTURE

Sharing of training data
- Upload
- Processing, Evaluation...
- SMT Resource Repository
- SMT Resource Directory

Training
- Giza++
- Moses SMT toolkit
- SMT Multi-Model Repository (trained SMT models)
- SMT System Directory
- Moses decoder

Using
- Anonymous access
- Web page
- Web page translation widget
- Web browser Plug-ins
- Authenticated access
- Web service
- CAT tools

System management, user authentication, access rights control...

Let's MT!
ACCURAT PROJECT
MISSION

To significantly improve **MT quality**

for **under-resourced** languages and narrow domains

by researching approaches how **comparable corpora** can compensate for a shortage of linguistic resources
COMPARABLE CORPORA

Non-parallel bi- or multilingual text resources

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
- containing overlapping information

Examples:
- multilingual news feeds,
- multilingual websites,
- Wikipedia articles,
- etc.
## Comparability Scale

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallel Corpora</td>
<td>- texts which are true and accurate translations;</td>
</tr>
<tr>
<td></td>
<td>- texts which are approximate translations;</td>
</tr>
<tr>
<td>Strongly Comparable</td>
<td>- texts from the same source on the same topic with the same editorial control;</td>
</tr>
<tr>
<td></td>
<td>- independently written texts on the same topic;</td>
</tr>
<tr>
<td>Weakly Comparable</td>
<td>- texts in the same narrow subject domain and genre;</td>
</tr>
<tr>
<td></td>
<td>- texts within the same broader domain and genre but varying in subdomains and specific genres;</td>
</tr>
<tr>
<td>Non-Comparable</td>
<td>- pairs of texts drawn at random from a pair of very large collections of texts (e.g. the web) in the two languages</td>
</tr>
</tbody>
</table>
KEY RESEARCH QUESTIONS

- How to measure comparability?
- How to collect comparable corpora?
- How to extract linguistic data for MT from comparable corpora?
- How to get most out of the data to improve SMT and RBMT?
- How to evaluate effect of our methods?
ACCURAT KEY OBJECTIVES

► To create comparability metrics - to develop the methodology and determine criteria to measure the comparability of source and target language documents in comparable corpora

► To develop, analyze and evaluate methods for automatic acquisition of comparable corpora from the Web

► To elaborate advanced techniques for extraction of lexical, terminological and other linguistic data from comparable corpora to provide training and customization data for MT

► To measure improvements from applying acquired data against baseline results from SMT and RBMT systems

► To evaluate and validate the ACCURAT project results in practical applications
ACCURAT LANGUAGES

► Focus on under-resourced languages
  Latvian, Lithuanian, Estonian, Greek, Croatian, Romanian, Slovenian

► Major translation directions
  e.g. English-Lithuanian. English-Croatian, German-Romanian

► Minor translation directions
  e.g. Lithuanian-Romanian, Romanian-Greek and Latvian-Lithuanian

► Methods will be adjustable to the new languages and domains and language independent where possible

► Applicability of methods will be evaluated in usage scenarios
ACCURAT PROJECT PARTNERS

► Tilde (Project Coordinator) - Latvia
► University of Sheffield - UK
► University of Leeds - UK
► Athena Research and Innovation Center in Information Communication and Knowledge Technologies - Greece
► University of Zagreb - Croatia
► DFKI - Germany
► Institute of Artificial Intelligence - Romania
► Linguatec - Germany
► Zemanta - Slovenia
EVALUATION OF EN-LV MT IN LOCALIZATION

► Goal: Increase in productivity of translators without degrading quality of translations
► Average increase of translators productivity: **32.9%**
► Increase of error rate from **20.2** to **28.6** points but still at the level “**GOOD**” (<30 points)
Web is becoming increasingly spoiled with low quality machine translated pages.

Tagging MT translated texts would help to avoid this data in MT training corpora.

Better domain/industry classification and related tags would help in collecting industry specific MT training data.

Common interfaces for MT engines would facilitate interoperability and integration in applications.
LET’S HELP SMALLER LANGUAGES TO BRIDGE TECHNOLOGICAL GAP!

letsmt.eu
accurat-project.eu
tilde.com
Andrejs Vasiljevs
andrejs@tilde.com