Enabling Users to Create Their Own Web-Based Machine Translation Engine

Andrejs Vasiļjevs
Tilde, Vienibas gatve 75a
Riga, Latvia
+371 67605001
andrejs@tilde.com

Raivis Skadiņš
Tilde, Vienibas gatve 75a
Riga, Latvia
+371 67605001
raivis.skadins@tilde.lv

Indra Sāmīte
Tilde, Vienibas gatve 75a
Riga, Latvia
+371 67605001
indra.samite@tilde.com

ABSTRACT
This paper presents European Union co-funded projects to advance the development and use of machine translation (MT) that will benefit from the possibilities provided by the Web. Current mass-market and online MT systems are of a general nature and perform poorly for smaller languages and domain specific texts. The ICT-PSP Programme project LetsMT! develops a user-driven machine translation "factory in the cloud" enabling web users to get customized MT that better fits their needs. Harnessing the huge potential of the web together with open statistical machine translation (SMT) technologies LetsMT! has created an innovative online collaborative platform for data sharing and building MT. Users can upload their parallel corpora to an online repository and generate user-tailored SMT systems based on user selected data. FP7 Programme project ACCURAT researches new methods for accumulating more data from the Web to improve the quality of data-driven machine translation systems. ACCURAT has created techniques and tools to use comparable corpora such as news feeds and multinational web pages. Although the majority of these texts are not direct translations, they share a lot of common paragraphs, sentences, phrases, terms and named entities in different languages which are useful for machine translation.

Categories and Subject Descriptors
H.4.m [Information Systems]: Miscellaneous

Keywords
LetsMT!, machine translation, Moses, data sharing, cloud service.

1. INTRODUCTION
As an exponential wave of information deluges the web, it becomes more important that access to information is available to people in their native languages. With 23 official languages the volume of information that requires translation in the European Union outpaces the ability of human translators to provide the service in a cost efficient manner. Machine translation is a critical tool for enabling European Union policies on language diversity and for communicating with citizens in their language.

Machine translation has been a particularly difficult problem in the area of natural language processing since its inception in the early 1940-ies. From the very beginning of MT history, three main MT strategies have been prominent: direct, interlingua, and transfer. Rule-based MT strategy with a rich translation lexicon showed good translation results and found its application in many commercial MT systems, e.g. Systran, PROMT and others. However, this strategy requires immense time and human resources to incorporate new language pairs or to enhance translation quality. The more competitive SMT approach has occupied the leading position since the first research results performed in the late 1980s with the Candide project at IBM for an English-to-French translation system [2]. The SMT strategy, first suggested in 1949 by Warren Weaver [23] and then abandoned for various philosophical and theoretical reasons for several decades until the late 1980s [3], has proven to be a fruitful approach to foster development of MT. Cost-effectiveness and translation quality are the key reasons that the SMT paradigm has become the dominant current framework for MT theory and practice [9]. However, these achievements do not fulfill all expectations regarding application of available SMT methods. The quality of an SMT system largely depends on the size of the training data. Obviously, the majority of parallel data is in the major languages. As a result SMT systems for larger languages are of much better quality compared to systems for under-resourced languages. This quality gap is further exacerbated by the complex linguistic structure of many smaller languages. Languages like Latvian, Lithuanian, and Estonian have a complex morphological structure and free word order. To learn this complexity from corpus data by statistical methods, much larger volumes of training data are needed than for languages with simpler structure. For example, Google Translator currently provides MT for more than 50 languages. However, for smaller languages quality is quite poor, particularly for domain specific texts. Another obstacle preventing wider use of MT is its general nature. Although free online translators provide reasonable quality for many language pairs, they perform poorly for domain and user-specific texts. Current online MT systems cannot be customized for particular terminology and style requirements or their adaptation is a prohibitively expensive service not affordable to smaller companies or public institutions.

The goal of the LetsMT! and ACCURAT projects is to overcome these challenges by taking advantage of the opportunities and multilingual data provided by the Web. These projects make state-of-art open source SMT tools easily accessible and unleash the huge potential of user-provided content to advance the quality and accessibility of machine translation.

2. LetsMT! OVERVIEW
LetsMT! is an online platform[1] that enables users to share translation data for MT training and to build tailored MT systems for different languages and domains on the basis of this data.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

[1] http://www.letsmt.com

The LetsMT! project is supported by the European Commission under the CIP ICT-PSP Programme. The LetsMT! Consortium includes project coordinator Tilde, the Universities of Edinburgh, Zagreb, Copenhagen and Uppsala, localization company Moravia and semantic technology company SemLab. The project was launched in March 2010 and will be completed by September 2012. The LetsMT! platform supports the following key features:

- Uploading of parallel texts for users that will contribute their content;
- Automated training of SMT engines from specified collections of training data;
- Custom building of MT engines from a selected pool of training data;
- Custom building of MT engines from proprietary non-public data;
- Storing and running user-generated MT engines;
- Automated MT evaluation.

LetsMT! translation services can be used in several ways: through the web portal, through a widget provided for free inclusion in a web-page, through browser plug-ins, and through integration in computer-assisted translation (CAT) tools and various online and offline applications. Localisation and translation businesses as well as other professional translators can use the LetsMT! platform to upload their parallel corpora in the LetsMT! website, build custom SMT solutions from selected collections of training data, and access these solutions in their productivity environments (typically, various CAT tools).

Figure 1 illustrates the general architecture of the LetsMT! platform. Its components for SMT training, parallel data collection and data processing are described further in this paper.

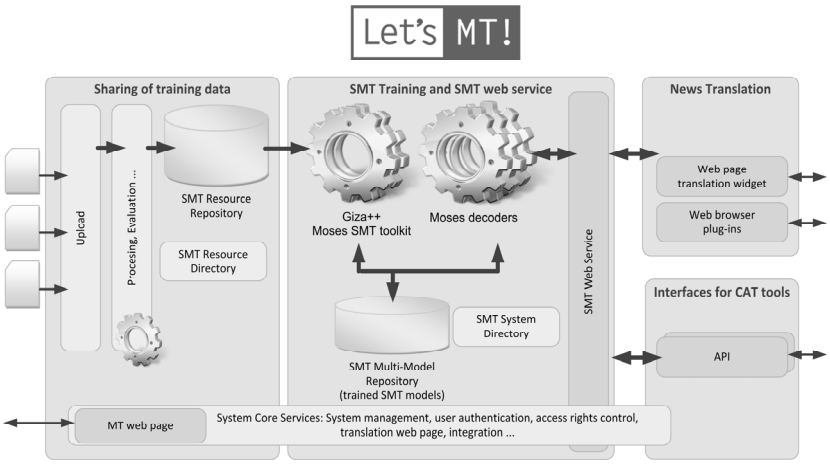

Figure 1. General architecture of the LetsMT! platform.

3. APPLICATION OF THE MOSES SMT TOOLKIT

A significant breakthrough in SMT was achieved by the EuroMatrix project. The project objectives included the creation of machine translation systems for all pairs of EU languages and the development of open source MT technology including research tools, software and data collections. Its result is the improved open source SMT toolkit Moses developed by the University of Edinburgh. The Moses SMT toolkit is a complete statistical translation system available under the Lesser General Public License (LGPL). Moses includes all the components needed to pre-process data and to train language and translation models [10]. Moses is widely used in the research community and has also penetrated the commercial sector.

LetsMT! uses Moses as a language independent SMT solution and integrates it as a cloud-based service into the LetsMT! On-line platform. One of the important achievements of the LetsMT! project is the adaptation of the Moses toolkit to fit into the rapid training, updating, and interactive access environment of the LetsMT! platform.

4. COLLECTING OF TRAINING DATA FROM THE WEB

While SMT tools are language independent, they require very large parallel corpora for training translation models. A parallel corpus is a collection of texts, each of which is translated into one or more languages [4]. SMT generates translations on the basis of statistical models with parameters derived from the analysis of bilingual parallel text corpora.

Thus, large scale parallel corpora are indispensable language resources for SMT [7]. Parallel corpora for smaller languages and domains are very limited in quantity, genre and language coverage. This remains true despite the creation of automated methods to collect parallel texts from the Web [7][8][11][1][14][15][18].

The most multilingual parallel corpus, the JRC-Acquis is a huge collection of European Union legislative documents translated into more than twenty official European languages [19] including under-resourced languages such as Latvian, Lithuanian, Estonian, Greek, Romanian, and others. For example, for the Latvian language it has 22,906 texts containing 27,592,514 words; for the Lithuanian language – 23,379 texts containing 26,937,773 words (version 3.0).

These resources along with other publicly available parallel resources, such as OPUS[2] [20] and JRC-Acquis[3] [19], are used in LetsMT! as initial training data for the development of pre-trained SMT systems. But there is a need for much more data in different domains. Translation systems trained on data from a particular domain, e.g. parliamentary proceedings, will perform poorly when used to translate texts from a different domain, e.g. news articles [16].

One of the solutions is to exploit the fact that comparable corpora, i.e., non-parallel bi- or multilingual text resources are much more widely available than parallel translation data. In contrast to parallel corpus, a comparable corpus can be defined as collection of similar documents that are collected according to a set of criteria, e.g. the same proportions of texts of the same genre in the same domain from the same period [12] in more than one language or variety of languages [4] that contain overlapping information [15] [8]. Comparable data like news feeds in different languages or multilingual webpages of international companies are produced every day and are available on the Web for many languages and domains.

The ACCURAT project has researched methods and developed tools to collect comparable corpora from the Web and to use this data to improve the quality of machine translation. The

[2] http://urd.let.rug.nl/tiedeman/OPUS/
[3] http://langtech.jrc.it/JRC-Acquis.html

ACCURAT toolkit[4] includes tools for corpus acquisition from the Web, comparability metrics allowing evaluations of the usability of collected corpora for MT tasks, and tools for multi-level alignment and extraction of lexical data for MT. These tools extract parallel sentences, phrases, terminology and named entities providing additional data for training statistical MT systems on LetsMT! or other platforms.

5. DATA PROCESSING

The LetsMT! project encourages the sharing and re-use of these valuable linguistic resources. The LetsMT! project provides opportunities for data holders, especially those in the public sector, to share their resources for the public benefit and for enabling citizens and users to get better quality machine translations of their content. The re-use of public sector information is mostly promoted through legislation that is binding upon Member States and European institutions (Directive 2003/98/EC on the re-use of public sector information and Commission Decision 2006/291/EC on the re-use of Commission information).

The motivation of users to get involved in sharing their resources is based on the following factors:

- an altruistic desire to participate and contribute, in a reciprocal manner, in a community of professionals and its goals;
- a way of building tailored and domain specific translation services;
- for individuals and businesses, a way to boost their reputations;
- for public institutions, a convenient means of ensuring compliance to the requirement set forth by EU Directive to ensure usability of public information;
- for academic institutions, a ready resource for study and teaching purposes.

To cope with the variety of user provided data, the project will resolve issues related to processing noisy data and ensuring data interoperability. The platform will discourage abuse and the inclusion of corrupted material, even though user authentication is used to reduce such dangers. The component of data management will therefore include various tests and pre-processing tools to validate the data and fix potential errors.

Besides basic validation, the LetsMT! platform requires a number of other pre-processing steps. Most important is a proper tokenisation module, since most of the users will not provide segmented data. Tokenisation is a non-trivial task and highly language dependent. In the first phase, simple standard tools are applied that split punctuations from other tokens, e.g. pattern-based tokenisation with the tools provided together with the Europarl parallel corpus. Language specific tools are used where they are available. Tools for better support of language specific issues will be continuously incorporated, like morphological analysers and lemmatisers.

Initially only user-provided translation memories containing aligned single-sentence units will be supported. Sanity checks should be carried out to avoid unreasonable training examples such as very long and fragmented translation units or sentences with formatting mark-up or other types of non-textual contents. At a later stage, LetsMT! will also support an upload of other types of parallel data like translations in pdf or doc format. The idea is to use existing resources in various formats and allow users to create their own training material in the form of sentence aligned corpora.

Support for number of the most used formats should be provided and the validation process ensured. Standard approaches to automatic sentence alignment are readily available, e.g. Hunalign [21], Vanilla [5], GMA [13]. Post-editing interfaces will be included to verify and improve alignment results online, e.g. ISA as part of Uplug. In this way, more users will be encouraged to provide parallel data in variety of formats.

The next step in building SMT translation models from parallel corpora is automatic word alignment. This part of the process is especially complicated and requires a great deal of computational power especially for large-scale corpora. Standard word alignment for SMT are the IBM models [3] and the HMM alignment model [22] implemented in the freely available tool GIZA++ [17].

Word alignment is time consuming and requires large amounts of internal memory for extensive data sets. Fortunately, there are extensions and alternative tools available with improved efficiency. LetsMT! has implemented the multi-threaded version of GIZA++ [6] that can run several word alignment processes in parallel on a multi-core engine.

6. APPLICATION SCENARIOS

LetsMT! puts a particular focus on elaborating two application scenarios, providing a detailed service concept for applications in the localization and translation industry and in free online translation of business and financial news.

The goal of the **online MT service for the localization and translation industry** is to increase the efficiency of localization and translation work performed by industry professionals – localization and translation service providers (LSPs), enterprises and organizations with multilingual translation needs, and freelance professionals of the language industry, through application of LetsMT! services that support generation of customized MT of higher quality, especially for smaller languages, and provide integration with professional productivity tools of the localization industry.

This service enables professional users to generate and employ customized MT services of higher quality based on specific terminology and style required by their clients. It takes into account the workflow, technical requirements and legal ramifications characteristic of the localization industry. The initial collection of corpora particularly focuses on parallel texts in smaller European languages like Latvian, Lithuanian, Estonian and Croatian. Once the service will come out of the beta stage, the range of domains and languages supported will be largely user-driven, i.e., determined by the requirements and opportunities in the localization and translation market.

The **online MT service for global business and financial news** enables users to follow the ever changing multilingual business news in their native languages. It will allow the public to use the LetsMT! translation engine for delivering "large language" news feeds to "small languages". The receiving clients will range from generic news readers to third party dissemination channels and proprietary news processing systems – anyone with an interest in reading or providing news stories from across the world in a variety of local languages. The online news translation service will be integrated in websites of business and financial news as well as analytical sites like newssentiment.eu.

[4] http://www.accurat-project.eu/index.php?p=toolkit

7. SYSTEM DEMONSTRATION

To demonstrate the capabilities of the LetsMT! platform we propose to have a live demo of the full cycle of an SMT engine generation. As training of a real SMT engine would require several millions of parallel sentences and would take several hours, we propose to use for the demo purpose a small dataset of limited vocabulary. We will show how in just about 10 minutes we can create a new MT system trained on this data. We will show how to use user-generated MT system for translation of web-pages, how to analyze multilingual financial news feeds and how to apply custom MT to increase efficiency in professional translation.

8. CONCLUSION

Current development of the SMT tools and techniques has reached the level that they can be implemented in practical applications addressing the needs of large user groups in variety of application scenarios. We invite Beta testers to visit LetsMT! website, use pre-trained MT systems, try MT customization features and contribute to MT development by uploading their parallel data. ACCURAT Toolkit can help to get more data from the Web for MT training and are provided as an open source tools to facilitate further elaboration. The current results of user evaluation indicate that these projects are developing in a direction that is demanded by potential users.

Successful implementation of the projects described in this paper will democratize access to custom MT, advance MT for under-resourced languages, and facilitate diversification of MT services by tailoring to specific domains and user requirements. These developments are example of tools and services needed to facilitate multilingual Web and reduce the language barriers.

9. ACKNOWLEDGMENTS

The research within the LetsMT! project has received funding from the ICT Policy Support Programme (ICT PSP), Theme 5 – Multilingual web, grant agreement 250456. The research within the project ACCURAT has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement 248347

10. REFERENCES

[1] Alegria, I., Ezeiza, N., Fernandez, I. 2008. Translating Named Entities using Comparable Corpora. Proceedings of the Workshop on Comparable Corpora, LREC'08, pp.11-17.

[2] Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, F., Roossin, P. 1988a. A statistical approach to French/English translation.

[3] Brown, P., Della Pietra, S., Della Pietra, V., Mercer, R. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics* 19.2 (19933): 264-311.

[4] EAGLES. 1996. Preliminary recommendations on corpus typology. Electronic resource: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html

[5] Gale, W.A. and Church, K.W. 1993. A Program for Aligning Sentences. *Bilingual Corpora. Computational Linguistics*, 1993, 19(1): 75- 102.

[6] Gao, Q. and Vogel, S. 2008. Parallel Implementations of Word Alignment Tool. In *Proceedings of Software Engineering, Testing, and Quality Assurance for Natural Language Processing*. pp. 49-57.

[7] Goutte, C., Cancedda, N., Dymetman, M., Foster, G. (eds.) 2009. *Learning Machine Translation*. The MIT Press. Cambridge, Massachusetts, London, England.

[8] Hewavitharana, S. and Vogel, S. 2008. Enhancing a Statistical Machine Translation System by using an Automatically Extracted Parallel Corpus from Comparable Sources. In *Proceedings of the Workshop on Comparable Corpora, LREC'08*, pp. 7-10.

[9] Hutchins, J. 2007. Machine translation: a concise history. In *Computer aided translation: Theory and practice*

[10] Koehn, P., Federico, M., Cowan, B., Zens, R., Duer, C., Bojar, O., Constantin, A., Herbst, E. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177-180, Prague

[11] Maia, B. and Matos, S. 2008. Corpógrafo V.4 – Tools for Researchers and Teachers Using Comparable Corpora. In Proceedings of the Workshop on Comparable Corpora, LREC'08, pp. 79-82.

[12] McEnery, A.M. and Xiao, R.Z. 2007. Parallel and comparable corpora: What are they up to? In *Incorporating Corpora: Translation and the Linguist. Translating Europe. Multilingual Matters*, Clevedon, UK.

[13] Melamed, D. 1999. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1), 1999, 107-130.

[14] Munteanu, D. 2006. Exploiting Comparable Corpora (for automatic creation of parallel corpora). Online presentation. Electronic resource: http://content.digitalwell.washington.edu/msr/external_release_talks_12_05_2005/14008/lecture.htm

[15] Munteanu, D. and Marcu, D. 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4): 477-504.

[16] Munteanu, D., Fraser, A., Marcu, D. 2004. Improved Machine Translation Performance via Parallel Sentence Extraction from Comparable Corpora. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT / NAACL'04*.

[17] Och, F.J and Ney, H. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, (29)1: 19-51.

[18] Resnik, P. and Smith, N. 2003. The Web as a Parallel Corpus. *Computational Linguistics*, 29(3) pp. 349-380.

[19] Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation: LREC'06*

[20] Tiedemann, J. 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. *Recent Advances in Natural Language Processing (vol V)*, John Benjamins, Amsterdam/Philadelphia, 237-248.

[21] Varga, D., Németh, L., Halįcsy, P., Kornai, A., Trón, V., Nagy, V. 2005. Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing*, pp. 590–596.

[22] Vogel, S., Ney, H., Tillmann, C. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, Denmark, 1996: (2).

[23] Weaver, W. 1949. Translation. Reprinted in Nirenburg, Somers and Wilks, Readings in Machine Translation, The MIT Press, 2003, pp. 13-17.