LetsMT! - Platform to Drive Development and Application of Statistical Machine Translation

Andrejs Vasiļjevs
Tilde
Vienibas gatve 75a, Riga, LV2101, LATVIA
E-mail: andrejs@tilde.lv

Abstract

This paper presents ICT-PSP project LetsMT! which develops a user-driven machine translation "factory on the cloud". Current mass-market and online MT systems are of general nature, system adaptation for specific needs is prohibitively expensive service not affordable to smaller companies or public institutions. To exploit the huge potential of open statistical machine translation (SMT) technologies LetsMT! has created an innovative online collaborative platform for data sharing and MT building.

Keywords: machine translation, cloud, SMT, parallel corpora, data processing

1. Introduction

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.

Another significant challenge is to break down the access barriers to SMT technology by making this platform and process user-friendly. Currently the implementation of SMT solutions, whether proprietary or out-sourced, requires an intensive investment of resources: human (natural language processing experts, system administrators), financial, and linguistic, to create and maintain a custom SMT infrastructure (Varga et al. 2005).

2. Project objectives

LetsMT! is a collaborative platform that thrives on resources contributed by its users. It contributes in a breakthrough 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.

LetsMT! provides a platform that supports the following features:

- Uploading of parallel texts for users that will contribute their own content;
- 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.

LetsMT! platform results from a project with duration of 30 months. It was started on March 1st 2010 and is planned to end by August, 2012. The project consortium consists of 6 partners – Tilde SIA, University of Edinburgh, University of Zagreb, University of Copenhagen, Uppsala University, Zoorobotics BV, Moravia. The project is coordinated by Tilde.

The core objective of the project is to provide innovative online MT services through sharing of parallel corpora provided by users, with emphasis on less-covered languages and specialized domains.

The solution created in the project provides the following core functions:

- A website for uploading parallel corpora and building specific MT solutions;
- A website for translation, where source text can be typed and translated;
- A 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. Platform and infrastructure

Work on the LetsMT! platform and infrastructure is the core activity within the project. The LetsMT! platform includes modules for sharing of SMT training data, SMT training and running, use in a news translation scenario, and use in a localisation usage scenario.

The beta versions of all the main modules is completed and deployed. The project Consortium has developed a common platform and supporting software infrastructure that provides the core functions necessary to integrate the modules of the LetsMT! platform. The supporting software infrastructure includes: the LetsMT! website, an API for external systems, User Management and Access Rights Control, Application Logic, an MT web page where users can try trained MT systems, etc.

It is obvious that hosting the LetsMT! platform requires a lot of computing capacity. The Project Consortium,

instead of buying servers, intends to lease capacity. It is economically efficient and will provide flexibility in adding new resources as necessary. During the analysis of detailed requirements, it was discovered that operating the LetsMT! platform on AWS (Amazon Web Services) was the most economically efficient option. It is planned to deploy the LetsMT! platform completely within the AWS, as this is a well-established solution. The AWS cloud provides a reliable and scalable infrastructure for deploying web-scale solutions. Alternative cloud computing suppliers may be selected if AWS fails to meet the requirements of the LetsMT! platform. The LetsMT! platform also can be deployed on a local server infrastructure.

4. SMT resource repository

The backend of the LetsMT! platform includes a modular resource repository. Figure 1 illustrates the general architecture of the software. Its design emphasizes possibilities of running the system in a distributed environment which makes the system suitable for scalable cloud-based solutions.

Communication between the web-frontend and the individual modules is handled by secure web service connections. A central database handles metadata information in a flexible key-value store that supports schema-less expandable information collections. The physical data storage can be distributed over several servers to reduce bottlenecks when transferring large data collections. Data collections can be stored using a version-controlled file system that supports data recovery and history management in a multi-user environment. The repository provides essential features for importing documents to the LetsMT! platform. Documents are converted, and sentences in translated documents are aligned automatically. The software is connected to a high-performance cluster that can execute various jobs with connection to the data stored in the repository, for example, the import and alignment of jobs. A cloud-based cluster enables scalability of the system according to the needs of the platform. The repository software is fully integrated in the current LetsMT! platform and can easily be extended with additional modules.

5. Collecting the training data

A large amount of training data is crucial for statistical machine translation.

The aim of the LetsMT! project is to collect data from both general language and from different subject domains. A special effort is being made by two of the project partners to collect business and finance news and localisation texts, mostly from the IT domain. Other subject domains are interesting for the project, so the partners focus on finding text providers with general language texts, in addition to domain specific texts.

The initial training corpora focused on Croatian, Czech, Danish, Dutch, Latvian, Lithuanian, Polish, Slovak, and Swedish. We still focus on these original languages, though other languages are also collected as part of multilingual corpora (Tiedemann 2009, Steinberger et al. 2006, Koehn 2005).

During the first year, lots of publicly available data was collected and provided on LetsMT! repository. Now Project Consortium concentrates on identifying new text providers and potential future users of the LetsMT! system.

For business and finance news, the Project Consortium uses a list of the largest companies from the involved countries to automatically harvest the newest parallel texts from these companies, and therefore, the collection is steadily growing.

The collection of parallel texts from the general language and from other subject domains is being advanced by making contacts at different levels. At the international level, the Project Consortium is in contact with TAUS (which has one of the largest repositories of parallel corpora) and with various EU institutions and projects, e.g., ACCURAT, TTC and META-NORD. At the national

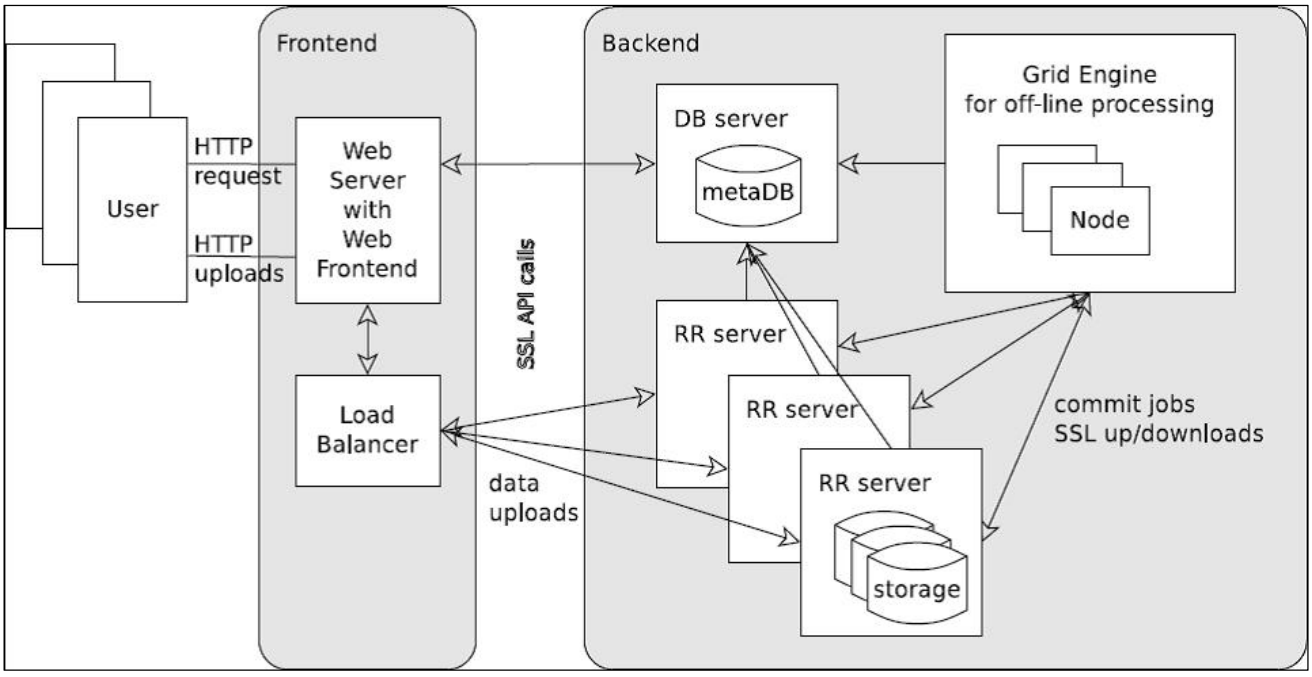

Figure 1: General architecture of a Resource Repository

level, project partner - Tilde has made a cooperation agreement with the National Library of Latvia. The partner University of Zagreb has made contact with several translation and localisation companies that are interested in the project and two of these have committed themselves to become text providers. The project partner-Moravia has made contact with the Slovak national corpus, but due to IPR problems their corpora cannot be used outside the institute. The project partner-Uppsala University contacted two institutes at Stockholm University who might be interested in using LetsMT!. The project partner University of Copenhagen contacted several potential text providers and has received acceptance from a company that write press releases in the EU languages and from at least 12 companies with annual reports. Furthermore, University of Copenhagen has started co-operation with a translation centre connected to University about texts in the domain of university administration.

The LetsMT! project has been presented at various events both at the national and international levels in order to spread knowledge and create awareness of the project in general and of the need for data in particular. Generally, the responses to the presentations are very positive, but IPR constitutes a challenge to the project. It turns out that some of the texts originally identified cannot be used outside the company/institution and thus cannot be uploaded to an external server like LetsMT!. Others can only be uploaded for private use and will not become publicly available on the LetsMT! platform. However, some of the contacts mentioned above are ready to sign a text provider agreement and others will follow soon.

6. SMT Training and running facilities

Users of the LetsMT! platform may select training resources from the SMT Resource Repository and train tailored SMT systems using the selected training resources. SMT training facilities include the following features: a user interface for resource selection and system training, integration with user authentication and access rights module, integration with SMT Resource Repository, simultaneous and effective execution of resources consuming training tasks, an interface providing information about running training tasks, progress, status, etc.

The SMT training facility and web service is built on top of the Moses machine translation toolkit (Koehn et al. 2007). Originally developed at the University of Edinburgh in 2007, Moses has since undergone a great deal of evolution. Many new features have been added, improving translation quality and keeping Moses up to date with the cutting edge of MT research. While of great importance, translation quality, however, is not the only aspect to have been worked on. SMT is extremely computationally demanding. Literally millions of options must be searched through in order to translate a single sentence, and the amount of data required to do so far outstrips the resources of and average desktop computer. Therefore, much research has been conducted on how to speed this process up and reduce the computational resources needed for translation.

Translation is only a part of what the Moses SMT Toolkit can do, though also included with it are the tools to train new translation systems. As with the actual method of translating, huge amounts of work have gone into training systems to yield better translations, as well as making the training process itself less resource intensive. The process of training a translation system is very in depth and intricate, but that too is handled by the toolkit.

Despite all the work that has gone into developing Moses, there are a few features required by the LetsMT! platform that Moses did not have. Having been conceived in academia, the focus of Moses has generally been towards features required by researchers and researchers. However, the environment in which it operates in the LetsMT! platform is very different. Developing Moses to support these new requirements is the main focuses of project activities and the work done in doing so is detailed below.

End users expect a service that delivers translations in a fast, interactive manner. Translating sentences requires a large amount of data, and waiting for this to be loaded each time would make the interactive user experience impossible to deliver. This has been addressed by the implementation of a version of Moses which runs on a background server, can be given sentences to translate interactively, and returns the translations quickly — without having to wait for the whole system to load up.

Users of the LetsMT! platform will also be translating between many different pairs of languages, and therefore, separate background processes for each pair would be impractical. This has been countered by allowing Moses to simultaneously have multiple translation systems in memory and by providing the language to translate into along with the sentence.

Modern computers are increasingly geared towards executing many processes in parallel, instead of doing them sequentially. In order to make the best use of available resources, Moses must be able to translate sentences in parallel. This feature, called 'multi-threading' has been integrated into Moses and enables it to deliver many translations in a fraction of the time compared to doing them one after another.

Other features such as being able to leverage new data without having to retrain the entire system have also been implemented and are in the process of being integrated with the rest of the platform. Methods for improving the fluency of translations using many billions of words of text are also in active development.

The LetsMT! platform is a great example of an EU project putting cutting edge technology to great use for the wider public, and as it does so, feeding back improvements to the academic community from where its ideas originate.

7. Usage scenarios

In particular, two specialized usage scenarios are supported by the LetsMT! platform: 1) machine translation of financial news, and 2) translation process in localization industry companies.

4.1 MT usage in news translation

Project Consortium has implemented the widget and browser (Mozilla, Internet Explorer) plug-ins of the LetsMT! platform.

The business scenario was developed in which the use of the widget is described. The aim for the business scenario is to provide translated business and financial information

through several facilities. There are two scenarios which are currently being investigated, a free and a paid, professional service. Free services will attract a broad audience of users with an interest in business related news and financial background information. The content will be information with a high latency, background information of local stock markets, local listed company information and comments. For low latency and emerging news, users can subscribe to a paid service. The targeted users are professionals and individuals that are interested in local and international breaking news and financial information. At the moment, the LetsMT! widget is integrated into SemLab's (Zoorobotics) business and financial news website www.newssentiment.eu for trial and evaluation purposes. The system is being tested on the website to ensure positive results in dissemination and exploitation activities through other (financial news) websites.

4.2 MT usage in localization

Professional users need MT services integrated in their working environment. Translators use CAT (Computer Aided Translation) tools (such as SDL Trados and MemoQ) in everyday activities. One of the prerequisites conditioning successful localisation scenario implementation is, without any doubt, the integration with CAT tools. In order to fulfil these requirements, the Project Consortium has developed a LetsMT! platform plug-in for SDL Trados Studio 2009 which allows for the use of the LetsMT! platform during translation process and experimentation on the evaluation of an English-Latvian SMT system applied to an actual localisation assignment (Vasiļjevs et al. 2011). The paper shows that such an integrated localisation environment can increase the productivity of localisation by 32.9% without critical reduction in quality.

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. The consortium partners are inviting Beta testers to evaluate the LetsMT! system and the current positive reviews on user experience indicate that the project is developing in a direction that is demanded by potential users.

Successful implementation of the project will democratize access to custom MT and, facilitate diversification of free MT by tailoring to specific domains and user requirements and have a strong impact on reducing the digital information divide in the EU.

9. Acknowledgements

The research within the LetsMT! project leading to these results has received funding from the ICT Policy Support Programme (ICT PSP), Theme 5 – Multilingual web, grant agreement no 250456. The research within the project ACCURAT has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no 248347.

10. References

Koehn, P. (2005). Europarl: a parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, Phuket, Thailand

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

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*

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.

Varga, D., Németh, L., Halļcsy, 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.

Vasiļjevs A., Skadiņš R. and Inguna Skadiņa I. 2011, Towards Application of User-Tailored Machine Translation, *Proceedings of Third Joint EM+/CNGL Workshop "Bringing MT to the User: Research Meets Translators" JEC 2011*, Luxembourg, 2011