LetsMT! – Online Platform for Sharing Training Data and Building User Tailored Machine Translation

Andrejs VASILJEVS, Tatiana GORNOSTAY and Raivis SKADINS
Tilde, Latvia

Abstract. This position paper presents the recently started European collaboration project LetsMT!. This project creates a platform that gathers public and user-provided MT training data and generates multiple MT systems by combining and prioritizing this data. The project extends the use of existing state-of-the-art SMT methods that are applied to data supplied by users to increase quality, scope and language coverage of machine translation. The paper describes the background and motivation for this work, key approaches, and the technologies used.

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

Introduction

In recent years statistical machine translation (SMT) has become a major breakthrough in machine translation (MT) development providing a cost effective and fast way to build MT systems. This development was particularly facilitated by the open-source corpus alignment tool GIZA++ [1] and the MT training and decoding tool Moses [2]. Another factor for facilitating the development of MT for many languages was the EU translation corpus and other parallel data available on the Internet. The EuroMatrix project has demonstrated how open source tools and publicly available data can be used to generate SMT systems for all language pairs of the official EU languages [3].

However, these achievements do not fulfil all expectations regarding the application of available SMT methods. The quality of an SMT system largely depends on the size of training data. Obviously, the majority of parallel data is in widely-used languages (e.g. English, German and some others). As a result, SMT systems for these languages are of much better quality compared to systems for under-resourced languages, i.e. languages with scarce linguistic resources. This quality gap is further deepened due to the complex linguistic structure of many smaller languages. Languages like Latvian, Lithuanian and Estonian, to name just a few, 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 the simpler linguistic structure.

Current systems are built on the data accessible on the web, but it is just a fraction of all parallel texts. The majority of valuable parallel texts still reside in the local systems of different corporations, public and private institutions, and desktops of individual users.

Another obstacle preventing wider use of MT is its general nature. Although free web translators provide reasonable quality for many language pairs, they perform poorly for domain and user-specific texts. Current free systems cannot be adjusted for particular terminology and style requirements. Large international corporations contract MT companies like Language Weaver to adapt translation systems for their particular needs. But this costly process is not accessible to smaller companies or the majority of public institutions. This prevents large part of the EU population from using existing MT solutions to get access to online information.

Specifically regarding application in the localization and translation industry, a huge number of parallel texts in a variety of industry formats have been accumulated, but the application of this data does not fully utilize the benefits of the modern MT technology. At the same time, this industry experiences a growing pressure on efficiency and performance, especially due to the fact that volumes of texts to be translated grow at a higher rate compared to availability of human translation, and translation results are expected in real-time. At present, the integration of MT in localization services is in its early stages, and the cost of developing specialized MT solutions is prohibitive to most players in the localization and translation industry. The quality of the generic MT offerings provided for free is too low to reap any efficiency gains in the professional localization industry setting. The same problem is faced by online information providers. They provide information mostly in the larger languages because the cost of human translation into smaller languages is prohibitively high and the quality of existing MT solutions is insufficient.

To fully exploit the huge potential of existing open SMT technologies and the huge potential of user-provided content, we propose to build an innovative online platform for data sharing and MT building. This platform is being created in the EU collaboration project LetsMT!. The LetsMT! Consortium includes the project coordinator Tilde, Universities of Edinburgh, Zagreb, Copenhagen and Uppsala, localization company Moravia and semantic technology company SemLab. The project started in March 2010 and should achieve its goals till September 2012.

The following sections describe the background and motivation of the LetsMT! project as well as the key approaches and technologies used.

1. Machine translation strategies

MT has been a particularly difficult problem in the area of natural language processing since its beginnings in the early 1940s. From the very beginning of MT history, three main MT strategies have been prominent: direct, interlingua, and transfer. The 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 were gained in the late 1980s with the Candide project at IBM for an English-to-French translation system [4][5]. The SMT strategy, first suggested in 1949 by Warren Weaver and then abandoned for various philosophical and theoretical reasons for several decades until the late 1980s [6], has proven to be a fruitful approach to foster the 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 [7].

In a majority of cases, SMT research and development activities were focused on widely used languages, such as English, German, French, Arabic, and Chinese. For smaller under-resourced languages, including the languages of the Baltic countries Lithuanian, Latvian and Estonian, MT solutions as well as language technologies in general have not been as well developed due to the lack of linguistic resources and cost effective technological approaches. This has resulted in a technological gap between these two groups of languages.

The goal of the LetsMT! project is to overcome this challenge by exploiting open source SMT toolkits and involving users in collecting training data. This will result in populating and enhancing the currently most progressive MT technology and making it available and accessible for all categories of users in the form of sharing MT training data and building tailored MT systems for different languages on the basis of the online LetsMT! platform.

2. LetsMT! approach

The LetsMT! project will extend the use of existing state-of-the-art SMT methods enabling users to participate in data collection and MT customization to increase quality, scope and language coverage of MT. Currently LetsMT! is creating a cloud-based platform that gathers public and user-provided MT training data and generates multiple MT systems by combining and prioritizing this data.

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

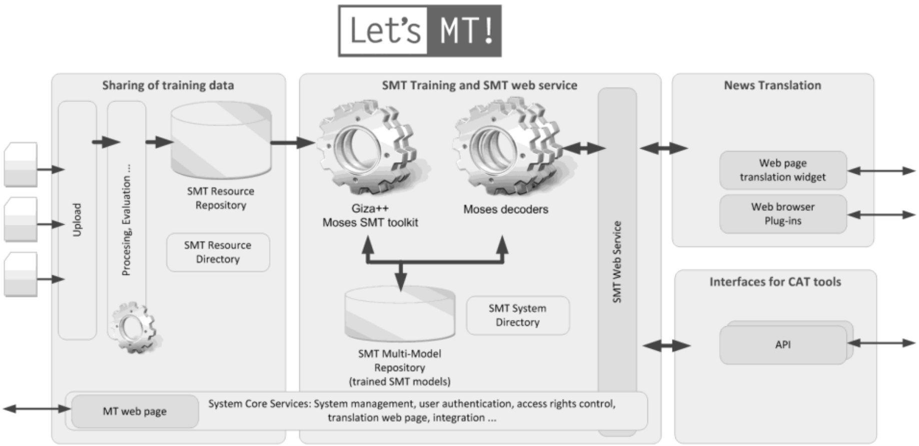
Let's MT!
Sharing of training data
Upload
Procesing, Evaluation ...
SMT Resource Repository
SMT Resource Directory
SMT Training and SMT web service
Giza++
Moses SMT toolkit
Moses decoders
SMT Multi-Model Repository (trained SMT models)
SMT System Directory
SMT Web Service
News Translation
Web page translation widget
Web browser Plug-ins
Interfaces for CAT tools
API
MT web page
System Core Services: System management, user authentication, access rights control, translation web page, integration ...

Figure 1. Software architecture of the LetsMT! platform.

LetsMT! services of translating texts will 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 different online and offline applications. Localisation and translation industry business and translation professionals will be able to use the LetsMT! platform for uploading their parallel corpora in the LetsMT! website, building custom SMT solutions from the specified collections of training data, and accessing these solutions in their productivity environments (typically, various CAT tools).

3. Application of the Moses SMT toolkit

A significant breakthrough in SMT was achieved by the EuroMatrix project[1]. Among project objectives were translation systems for all pairs of EU languages and the provision of the open source MT technology including research tools, software and data. Its result is the improved open source SMT toolkit Moses developed by the University of Edinburgh. The Moses SMT toolkit is a complete translation system distributed under the Lesser General Public License (LGPL). Moses includes all the components needed to pre-process data, train language models, and translation models [2]. Moses is widely used in the research community and has also reached the commercial sector. While the use of the software is not closely monitored (there is no need to sign a license agreement), Moses is known to be in commercial use by companies such as Systran, Asia Online, Autodesk, Matrixware, Translated.net. The LetsMT! project coordinator Tilde bases its free online Latvian MT system[2] on the Moses platform.

LetsMT! uses Moses as a language independent SMT solution and integrates it as a cloud-based service into the LetsMT! online platform. One of the important advancements of the LetsMT! project will be the adaptation of the Moses toolkit to fit into the rapid training, updating, and interactive access environment of the LetsMT! platform. The SMT training pipeline implemented in Moses currently involves a number of steps that each require a separate program to run. In the framework of LetsMT! this process will be streamlined and made automatically configurable given a set of user-specified variables (training corpora, language model data, dictionaries, tuning sets).

Additional important improvements of Moses that are being implemented by the University of Edinburgh as part of LetsMT!, are the incremental training of MT models, randomised language models [8], and separate language and translation model servers. We expect some users to add relatively small amounts of additional training data in frequent intervals. The incremental training will benefit from the addition of these data without re-running the entire training pipeline from scratch.

4. Parallel corpora for SMT training

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 [9]. 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 [10]. The most multilingual parallel corpus, the JRC-Acquis is a huge collection of European Union legislative documents translated into more than

[1] http://www.euromatrix.net/

[2] http://translate.tilde.com

twenty official European languages [11] 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[3]).

A similar corpus to JRC-Acquis is the European Parliament Proceedings Parallel Corpus[4] (Europarl corpus) which was extracted from the proceedings of the European Parliament (1996-2006) and includes versions in 11 European languages: French, Italian, Spanish, Portuguese, English, Dutch, German, Danish, Swedish, Greek and Finnish [12].

These resources along with other publicly available parallel resources, such as OPUS[5] and JRC-Acquis[6], are used in LetsMT! as initial training data for the development of pre-trained SMT systems.

4.1. Applying user-provided data for SMT training

The number of open source parallel resources is limited and this is an essential problem for SMT, since 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 [13][14]. At the same time, a huge amount of parallel texts and translated documents are at the users' disposal and they can be used for SMT system training. Therefore, the LetsMT! online platform will provide all categories of users (public organizations, private companies, individuals) with an opportunity to upload their proprietary resources to the repository and receive a tailored SMT system trained on these resources. The latter can be shared with other users who can exploit them further on.

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

- participate and contribute, in a reciprocal manner, with a community of professionals and its goals;
- achieve better MT quality for user specific texts;
- build tailored and domain specific translation services;
- enhance reputation for individuals and businesses;
- ensure compliance with the requirement set forth by EU Directive to provide usability of public information in a convenient way for public institutions;
- deliver a ready resource for study and teaching purposes for academic institutions.

The LetsMT! project is advancing the concept of data sharing which implies the practice of making data used in one activity available to other users. One of the examples of successful data sharing is the Translation Automation User Society (TAUS) Translation Memory (TM) Sharing Platform – TAUS Data Association (TDA)[7]. TDA is a global not-for-profit organization providing a neutral and secure platform for sharing language data. By sharing their TMs and glossaries, members in return get access to the data of all other members.

[3] http://langtech.jrc.it/JRC-Acquis.html
[4] http://www.statmt.org/europarl/
[5] http://urd.let.rug.nl/tiedeman/OPUS/
[6] http://langtech.jrc.it/JRC-Acquis.html
[7] http://www.tausdata.org

There is an obvious cooperation potential between LetsMT! and TDA. LetsMT! is interested in using TDA data for SMT training and TDA is interested in integrating MT generation capabilities with the TDA data repository. TDA is already a member of the LetsMT! Support Group and cooperation is further ensured by membership of project partners Tilde and Moravia in TDA.

4.2. Processing of training data

Since user-provided shared data plays a major role in LetsMT!, the project should deal with issues related to processing of noisy data and ensuring data interoperability. The platform should prevent its abuse by 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. There are various ready-made tools that can be used out-of-the-box for data checking. For example, freely available XML parsers, e.g. Libxml2[8], Tidy[9], and TMX[10] validators will be used to detect problems in the TMs provided by users; open-source GNU/Unix tools will be used to detect character encodings and perform conversions; language guessers, e.g. TextCat[11], are also available and can easily be trained for additional language/character sets [15]. It is feasible to use existing tools and integrate them in the LetsMT! platform in order to detect and correct potential errors.

Besides basic validation, the LetsMT! platform requires a number of other pre-processing steps. Most important is a proper tokenisation (text segmentation) 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 will be 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 will be 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. 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 a number of common formats should be provided and the validation process ensured. Standard approaches to automatic sentence alignment are readily available, e.g. Hunalign[12] [16], Vanilla[13] [17], GMA[14] [18]. Post-editing interfaces will be included to verify and improve alignment results online, e.g. ISA as part of

[8] http://xmlsoft.org/

[9] http://tidy.sourceforge.net/

[10] http://www.maxprograms.com/products/tmxvalidator.html

[11] http://www.let.rug.nl/vannoord/TextCat/

[12] http://mokk.bme.hu/resources/hunalign

[13] http://nl.ijs.si/telri/Vanilla/

[14] http://nlp.cs.nyu.edu/GMA

Uplug[15] [19]. In this way, more users will be encouraged to provide parallel data in a 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 [6] and the HMM alignment model [20] implemented in the freely available tool GIZA++ [1]. It can be used as a black-box tool in connection with the Moses toolkit which supports all the necessary steps to build a phrase-based SMT model from a given sentence aligned parallel corpus. Word alignment is carried out in an unsupervised way using EM re-estimation procedures and a cascaded combination of alignment models. Various settings can be adjusted in the alignment procedure and phrase table extraction.

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. Multi-threaded version of GIZA++[16] [21] can run several word alignment processes in parallel on a multi-core engine. Furthermore, another version of GIZA++ (cluster-based) can be used to distribute word alignment over various machines. An alternative approach that can also run a parallel alignment procedure is implemented in the MTTK toolkit[17] [22]. This software provides several alignment models and may also be used to perform sentence alignment which usually is a prerequisite for word alignment.

5. Conclusion

Current development of SMT tools and techniques has reached the level where they can be implemented in practical applications addressing the needs of large user groups in a variety of application scenarios. The work in progress that is described in this paper promises important advances in the application of SMT by integrating available tools and technologies into an easy-to-use cloud-based platform for data sharing and generation of customized MT.

Successful implementation of the project will enable wider use and greater impact of available open-source SMT technologies, facilitate diversification of free MT by tailoring it to specific domains and user requirements.

Acknowledgements

We would like to thank Jörg Tiedemann and other LetsMT! project partners for contributing to parts of this paper.

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 n°250456.

[15] http://www.let.rug.nl/~tiedeman/Uplug/php/

[16] http://code.google.com/p/giza-pp/ and http://www.cs.cmu.edu/~qing/

[17] http://mi.eng.cam.ac.uk/~wjb31/distrib/mttkv1/

References

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

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

[3] P. Koehn, A. Birch and R. Steinberger, 462 Machine Translation Systems for Europe, *in Proceedings of MT Summit XII*, 2009.

[4] P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, F. Mercer, P. Roossin, A statistical approach to French/English translation, *in Proceedings of the Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, June 12-14, 1988.

[5] P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, F. Mercer, P. Roossin, A statistical approach to language translation, *in Proceedings of the 12th International Conference on Computational Linguistics COLING'88*, (1): 71-76, 1988.

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

[7] J. Hutchins, Machine translation: a concise history, *in Computer aided translation: Theory and practice*, ed. Chan Sin Wai. Chinese University of Hong Kong, 2007.

[8] A. Levenberg, M. Osborne, Stream-based Randomised Language Models for SMT, *in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 2009.

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

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

[11] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiș, D. Varga, 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*. Electronic resource: http://langtech.jrc.it/Documents/0605_LREC_JRC-Acquis_Steinberger-et-al.pdf, 2006.

[12] P. Koehn, Europarl: a parallel corpus for statistical machine translation, *in Proceedings of Machine Translation Summit X*, 2005.

[13] D. Munteanu, A. Fraser, D. Marcu, 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*, Electronic resource: http://www.mt-archive.info/HLT-NAACL-2004-Munteanu.pdf. 2004.

[14] D. Munteanu, 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, 2006.

[15] W.B. Cavnar, J.M. Trenkle, N-Gram-Based Text Categorization, *in Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV, UNLV Publications/Reprographics, pp. 161-175, 11-13 April, 1994.

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

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

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

[19] J. Tiedemann, ISA & ICA - Two Web Interfaces for Interactive Alignment of Bitexts, *in Proceedings of LREC 2006*, Genova, Italy, 2006.

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

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

[22] Y. Deng, W. Byrne, MTTK: An alignment toolkit for statistical machine translation, in Demo Presentation in the HLT-NAACL Demonstrations Program, June 2006.