Machine translation for e-Government – the Baltic case

Andrejs Vasiļjevs andrejs@tilde.lv
Rihards Kalniņš rihards.kalnins@tilde.lv
Mārcis Pinnis marcis.pinnis@tilde.lv
Raivis Skadiņš raivis.skadins@tilde.lv
Tilde, Riga, LV-1004, Latvia

Abstract

This paper presents a case study about the development of MT systems for two Baltic governments. The governments of Latvia and Lithuania presented Tilde with a need to expand their communication to reach multilingual citizens. In order to meet this need, Tilde collected a vast amount of domain-specific data and trained MT system to produce high-quality translation. In the process, Tilde identified and overcame the challenges of a lack of parallel corporate for the given domain and language pairs. In this paper, we discuss how the systems were integrated into several facilities: a public online interface, website translator, webpage widgets, and mobile apps. We will detail how these facilities allow for the MT system to be used in various applications, including document translation, website translation, and integration into e-services. As a result of MT application, the public sector can not only make its services more universally acceptable, but also improve the flow of information to and from citizens.

1. Introduction

A critical function of government is to ensure access to services for all residents and visitors. A significant component of providing access to services is making them available in multiple languages. The ability to convey timely and necessary information across language barriers is of paramount importance in a multilingual environment.

In this paper we illustrate how two countries are implementing machine translation to reach out to citizens and make e-government services universally accessible.

Pioneers among European Union member states, both Latvia and Lithuania have embarked on targeted programs to build language technologies and custom machine translation into key e-government services. Having noted the potential and motivated by the requirements for multilingual language support within the EU, these Baltic countries are seeking language technology solutions for their multilingual communication needs.

In order to reach out to non-Latvian speaking residents and improve the flow of information to and from residents and visitors, the Latvian government has decided to pilot machine translation for the Latvian/Russian and Latvian/English language pairs. The objective of the project is to create high-quality custom machine translation systems for a number of domains and provide facilities for MT integration into e-government applications.

The specific challenges are a lack of parallel resources for the required domains and language pairs, as well as the complexity of the languages – specifically, Russian and Latvian are highly inflected languages with relatively free word order.

The Latvian e-government solution is based on the LetsMT (Vasiļjevs et al., 2012) platform provided by Tilde. LetsMT extends Moses technologies (Koehn et al., 2007) for the demands of dynamically scalable real-time applications. The solution supports multiple input formats, maintain tag and formatting integrity for translating documents. Integration in services is enabled through API, translation widgets, and browser add-ons.

The Lithuanian government is taking a similar approach, and has also selected to base its MT services on the LetsMT technologies. The Lithuanian projects have elected to create MT systems for Lithuanian/French and Lithuanian/English language pairs. A number of challenges are similar – such as a lack of resources, linguistic complexity, as well as additional requirement to provide MT service on mobile apps.

In presenting these two use cases, we detail technological, infrastructure, and linguistic challenges we have experienced and overcome while creating and piloting these large scale public MT projects. We also describe a number of innovations that have been tested and applied to the creation of the machine translation engines to compensate for the lack of parallel data and to boost translation quality, e.g. such as terminology integration in SMT and training data extraction from comparable corpora.

We discuss the application of the developed MT systems, in particular to facilitate Latvia's presidency of the Council of the European Union. We also discuss the role of these pilots in the implementation of the envisioned EU public automated translation service infrastructure as part of the Connecting Europe Facilities Programme.

2. Motivation

The public sector in the European Union faces a variety of daunting tasks today. In addition to maintaining the daily activities of government and providing public services, the European public sector must also constantly engage and interact with a wide variety of citizens, informing them about government activities and available services.

Communication for the public sector can also be problematic, considering the wide range of languages spoken by European citizens. This is a particular challenge in a small country like Latvia, with a large minority group speaking a language different from the official language. Though the official language is Latvian, the 2011 census shows that approximately 37% of the Latvian population speaks Russian at home.[1] Therefore the provision of public services in the official language can lead to communication barriers.

This barrier has traditionally been overcome by the use of human translation. 71% of the country's ethnic Latvians also speak Russian, therefore the pool of potential LV-RUS translators is considerable. Human translation, however, has definite limitations in the public sector. The large number of texts that must be translated, as well as the speed in which translations are expected, can be overwhelming. In addition, human translation is very costly; this is a particular difficulty in Latvia and Lithuania, which rank among countries with the lowest GDP per capita in the EU.[2]

The Latvian public sector faces an additional challenge in communication: the country's small size. With less than 2 million native speakers, Latvian is one of the smallest languages in the European Union. This can lead to information isolation, as very few individuals outside of Latvia can speak the language fluently, therefore relying exclusively on a limited pool of translators. Latvia's accession to the EU in 2004 and the Eurozone in 2013 – and the increasingly

[1] http://www.csb.gov.lv/en/notikumi/home-latvian-spoken-62-latvian-population-majority-vidzeme-and-lubana-county-39158.html

[2] http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/GDP_per_capita,_consumption_per_capita_and_price_level_indices

larger role it is playing in EU affairs, such as hosting the Presidency of the Council of the European Union in 2015 – have also made the need for seamless communication between Latvia and the rest of Europe more pressing than ever.

In order to meet these challenges and ensure more efficient communication, the Latvian public sector commissioned Tilde to build a machine translation system. The goal of the MT system is to facilitate communication between the Latvian public sector and Latvia's citizens, as well as with the rest of the European Union. The MT system would overcome the need for human translation – which is both costly and slow – and would be specially adapted to the various needs of the public sector.

3. Requirements

The Latvian public sector presented Tilde with a list of requirements that the MT system should meet, so that it would be adapted to the needs of e-government. These included training the system on specific terms for legislative acts, public services, and e-gov content; and adapting the system to the functional needs of the public sector, such as integration into e-gov systems and services.

Given that the public sector covers a wide range of different tasks and shares a myriad of responsibilities, the MT system had to be capable of translating texts in a variety of domains: from public service information and enterprise registry data to museum databases and library catalogues. This would ensure that the MT system could be used by institutions as varied as the Enterprise Registry, the State Chancellery, the Ministry of Foreign Affairs, the National Museum of Art, and the National Library.

Tilde was also given a list of several use scenarios in which the MT system would be deployed. These included:

- widget in government websites, so web administrators can machine translate content in CMS and then perform light in-place post-editing, and visitors can machine translate Latvian-language content into Russian or English
- translation API, so that ministry programmers can choose where and how to include machine translation function in government sites
- website translator, so that government employees can translate websites using systems tailor-made for public sector content (e.g., legislation).

The most substantial use scenario was an online public MT interface – a website accessible from any computer connected to the Internet. On the site, visitors could machine translate phrases and sentences. They would also have the option to upload and translate entire documents in several formats: DOCV, PDF, or PPT. The system would be open to any visitors – either from Latvia or abroad. But it would have particular benefit for employees in the Latvian public sector, who can use the interface to translate documents, e-mails, PowerPoint presentations, or a range of other texts.

Due to the specific needs of the public sector, however, all information entered on the site would be translated securely, guaranteeing the confidentiality of sensitive data. The MT system would be hosted in an on-premise appliance at a public sector institution, further reinforcing the security of data. This means that the interface can also be used to translate highly confidential public sector documents, such as diplomatic communiques, Ministry of Defense documents, draft legislation, etc.

MT adaptation by the public sector presents its own challenges, however, such as domain adaptation and small language quality. For the public sector, MT quality requirements are often high. The public sector's specific needs for MT solutions have not been satisfied by current MT solutions. For instance, the public sector has a need to adjust MT systems to particular

language pairs and domains, e.g. specific topic, language style preferences, terminology, taxonomies, and named entities. Another important requirement is more widespread availability of MT: the public sector has a need for instant text translation, document translation, web page content translation, and MT integration in information systems and work processes. In order to meet these needs, a MT solution must be stable and highly scalable to process a high amount of translation tasks. Finally, the public sector has a need for highly secure MT solutions that could be used for translation of even classified information.

4. Solution architecture

Both Latvian and Lithuanian e-gov solutions are developed on the basis of the LetsMT! platform. This technology has resulted from the EU ICT-PSP supported research and industry collaboration project LetsMT!. To meet the set requirements initial platform was significantly extended with additional functionality and new modules were created.

SMT training and decoding facilities are based on the open source toolkit Moses. Moses includes the essential components needed to pre-process data and to train language and translation models. SMT training is automated using Moses experiment management system (Koehn, 2010).

The resulting solution has a multitier architecture inherited from the LetsMT platform (Figure 1). It has (i) an interface layer implementing the user interface and APIs with external systems; (ii) an application logic layer for the system logic, (iii) a data storage layer consisting of file and database storage and (iv) a high performance computing (HPC) cluster. The LetsMT! system is performing various time and resource consuming tasks; these tasks are defined by the application logic and the data storage and are sent to the HPC cluster for execution.

The interface layer provides interfaces between the system and external users - both human users and machine users like websites and third party solutions. Human users can access the system through web browsers by using web page interface. External systems such as web browser plug-ins can access MT services through a public API. The public API is available through both REST/JSON and SOAP protocol web services. An HTTPS protocol is used to ensure secure user authentication and secure data transfer.

The application logic layer contains a set of modules responsible for the main functionality or logic of the system. It receives queries and commands from the interface layer and prepares answers or performs tasks using the data storage and the HPC cluster. This layer contains several modules such as the Resource Repository Adapter, the User Manager, the SMT Training Manager etc. The interface layer accesses the application logic layer through both REST/JSON and SOAP protocol web services. The same protocols are used for communication between modules in the application logic layer.

The data is stored in one central Resource Repository (RR). As training data may change (for example, grow), the RR is based on a version-controlled file system (currently we use SVN as the backend system). A key-value store is used to keep metadata and statistics about training data and trained SMT systems. Modules from the application logic layer and HPC cluster access RR through a REST-based web service interface.

A High Performance Computing Cluster is used to execute many different computationally heavy data processing tasks – SMT training and running, corpora processing and converting etc. Modules from the application logic and data storage layers create jobs and send them to HPC cluster to execute. HPC cluster is responsible for accepting, scheduling, dispatching, and managing the remote and distributed execution of large numbers of standalone, parallel or interactive jobs. It also manages and schedules the allocation of distributed resources such as

processors, memory and disk space. The LetsMT! HPC cluster is based on the Oracle Grid Engine (SGE).

The hardware infrastructure of the LetsMT! platform is heterogeneous. The majority of services run on Linux platforms (Moses, RR, data processing tools etc.). The Web server and application logic services run on a Microsoft Windows platform.

The system hardware architecture is designed to be highly scalable. The system contains several machines with both continuous and on-demand availability. Continuous availability machines are used to run the core frontend and backend services, and HPC grid master that guarantee stable system functioning. On-demand availability machines are used (i) to scale up the system by adding more computing power to training, translation and data import services (HPC cluster nodes), (ii) to increase performance of frontend and backend server instances.

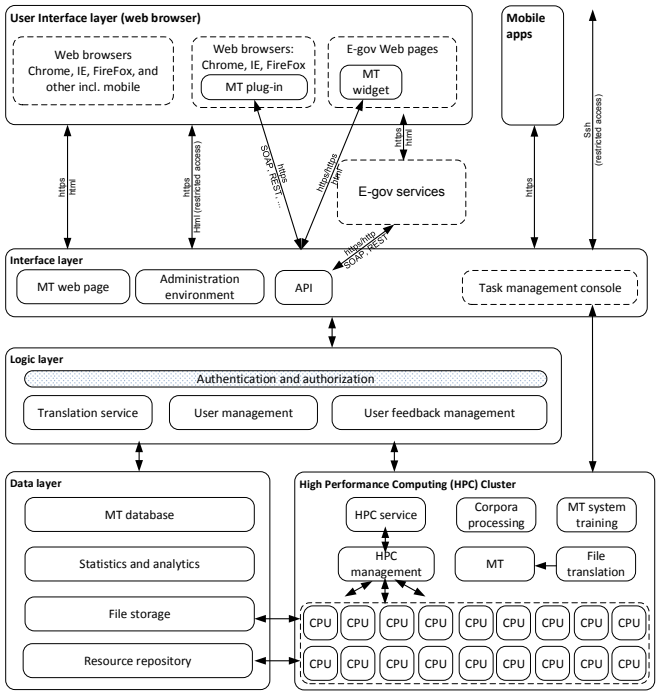

Figure 1. System architecture

To ensure secure and confidential processing of data solution is hosted on the local infrastructures of the client institutions.

Both Latvian and Lithuanian languages that are the core languages of the described projects are highly inflected. Combined with relatively scarce parallel corpora available for these languages this poses a major challenge for statistical MT.

To address this challenge we apply factored translation technique (Koehn and Hoang, 2007; Bojar, 2009) devised for morphologically rich languages. Standard SMT systems treat each word form as an individual unit, regarding different inflected forms of the same word as completely unrelated tokens, leading to high data sparseness for morphologically rich languages. Factored models try to alleviate this issue by analyzing the data with part-of-speech taggers and annotating each word form with its lemma and part-of-speech tag. This additional information is then provided to the SMT system, allowing it to draw better generalizations over

the data. Such approaches may be used both for the translation model and for the language model, as appropriate.

5. Data collection

To improve quality of MT systems particular efforts were devoted to the data collection in the required domain. The small size of the language and the country logically leads to a smaller set of data than might be found for larger languages elsewhere. As part of the LetsMT project, Tilde had already built a substantial Latvian-English-Latvian system, which has proven to perform better than Google Translate for the given language pair.

But the Latvian public sector's need for a domain-specific system called for addition data collection. To this end, Tilde relied on existing translations of legislative acts (parallel texts), ministry websites, and descriptions of e-government services. The lack of sufficient parallel data for the required level of quality, however, was a challenge. To this end, Tilde used a number of innovations. For data collection Tilde applied some new techniques to process comparable multilingual corpora. We used approached and tools developed at the ACCURAT project to extract data from multilingual news sites, Wikipedia, and other comparable text collections (Skadiņa et al., 2012).

For the needs of the project Lithuanian Parliament provided access to several thousand working documents with their translations in the DOC format. These documents were aligned with the Bilingual Sentence Aligner by Microsoft Research (Moore, 2002) producing 815 thousand parallel segments for Lithuanian-Russian and 235 thousand segments for Lithuanian-English. This data is used to customize MT for the needs of the Parliament Translation Department.

To increase amount of the data for customization of Latvian MT systems we reviewed the catalog of Latvian legislative acts and identified national adaptations and translations of international agreements and conventions. Pairing these documents with the corresponding counterparts found on the Web resulted in 619 Latvian-English and 427 Latvian-Russian parallel pairs of documents. This significantly boosted data needed to adapt MT for the use in legislative context.

The systems can be further improved by uploading new collections of organization's data – parallel corpora or human-translated documents. Solution includes facilities that automatically extract and align sentences from user uploaded parallel documents in the most popular formats such us PDF, RTF, DOC, DOCX, TMX, XLIFF.

6. Domain adaptation

We use the term "domain" here to broadly refer to a particular type of text, including style, genre (e.g. news versus email), and topic (finances versus legislation), but also a specific "brand" (e.g., EU legislation vs national legislation). Adapting for a domain thus results in a system that is better at translating a specific text style, type, topic, and brand. In that respect, the success of SMT hinges on the availability of the right data: when the text on which the model is trained closely matches the testing data in terms of style, genre and topic, the system outputs are often very good. Conversely, when there is a data mismatch, the performance degrades substantially (e.g., Foster et al. (2010) report a massive drop of 12% BLEU absolute for out- vs in-domain training between parliamentary speech and medical texts). Currently one of the most active areas of research in MT and language processing in general (e.g., Daumé (2007) and Blitzer et al. (2007)) is domain adaptation, which seeks answers for the question "Given large corpora of general training resources and a small corpus of in-domain data, how can these be used to obtain accurate in-domain translation?"

The majority of domain adaptation research has investigated methods for adapting the two core components of the SMT system, the language model and the translation model. These approaches estimate separate models on each domain, which are then combined using interpolation or backoff in order to maximise in-domain translation performance, which has been shown to be highly successful (Koehn and Schroeder, 2007; Nakov, 2008; Foster et al., 2010; Sennrich, 2012). Interestingly, many of these papers have shown that pure in-domain training is a very difficult baseline to beat, requiring careful tuning to ensure that the out-of-domain data does not swamp the in-domain data. Further gains can be had by augmenting the in-domain training data by selecting or re-weighting similar text in large out-of-domain monolingual or bilingual text collections (Zhao et al., 2004; Munteanu and Marcu, 2005; Matsoukas et al., 2009; Axelrod et al., 2011).

In the QT Public project we will build upon the latest research achieved by the TaaS project, adapt the methods for translation of domain-specific items and integrate the methods in framework for online SMT system adaptation.

7. Improving terminology translation

Documents and texts covered by e-gov services can be on many different subjects (from broad domains, e.g., economics, law, information and communication technologies, to numerous different narrow domains, e.g., diplomatic communication, medicine descriptions, etc.). The fact that the scope of e-government document types is so wide makes it very difficult for SMT systems to ensure good terminology translation quality. Each of the different domains also has its own typical terminology. Because there is a very high level of cross-domain ambiguity for e-gov text translation, typical SMT system adaptation methods that require re-training of SMT systems may not be feasible (simply because of uneconomical reasons). A possible solution is to perform dynamic terminology integration in SMT systems during document (or text) translation.

Related work has previously tried to address this challenge. For instance, Carl & Langlais (2002) in their research showed that using terminology dictionaries in order to identify terms in the translatable content and to provide translation equivalents from the dictionaries could increase the translation performance for the English-French language pair. Babych & Hartley (2003) showed that for NEs (namely, organization names) special "*do-not-translate*" lists allowed increasing translation quality for the English-Russian language pair using a similar pre-processing technique that restricts translation of identified phrases. However, these methods have been elaborated mostly for languages with simple morphology or categories of phrases that are rarely translated or even left untranslated (e.g., many company and organization names). A recent study in the FP7 project TTC (2013) has shown that for English-Latvian such simplified pre-processing does not yield positive results for term translation. Hálek et al. (2011) also showed that the translation performance with on-line pre-processing drops according to BLEU (Papineni et al., 2002) for English-Czech named entity translation. This proves that a more sophisticated method is necessary when translating into morphologically rich languages, or languages with high level of inflection (e.g., the Baltic and Slavic languages).

In order to address the challenges put forth by the cross-domain ambiguity of e-gov services and to provide effective means for terminology integration in SMT systems in order to ensure correct and consistent translation of terminology, we have designed a workflow for dynamic terminology integration in SMT systems (TaaS, 2014). The workflow utilizes a pre-trained SMT system and performs translatable content pre-processing prior to translation of the content with an SMT decoder. The translatable content pre-processing workflow is depicted in

Figure 2. It uses a bilingual term collection specified by the user (translator, e-gov service provider, etc.) in order to identify terms in the translatable content (e.g., a sentence, a paragraph, even a full document) using a shallow (but very efficient) method, annotates the content with possible translation hypotheses from the bilingual term collection using XML mark-up[3] that complies with the Moses SMT system's XML mark-up format, assigns translation confidence scores for each of the translation hypotheses, and, finally, translates the document taking into account the injected XML mark-up with the Moses SMT decoder.

In order to ensure that terms in different surface forms (or inflected forms) for morphologically rich languages are also identified, stemming is applied both to terms and phrases in the translatable content. As translation speed is very critical for production systems, such a shallow approach is a compromise between speed and quality.

Terms, depending on the type (or source) of a term collection, can be given in their canonical forms (as stored in dictionaries) or in different surface forms. If the terms are stored in canonical forms, for morphologically rich languages it is necessary to acquire or generate all possible surface forms that can be potential translations of the terms. In the proposed workflow, this task is handled by the term translation acquisition process. Practically, there are two possible methods: 1) term surface forms can be looked-up in a large monolingual corpus in the target language, or 2) term surface forms can be generated using morpho-syntactic transformation rules. The first method is language independent (if we ignore the fact that stemming tools are necessary in order to identify the different surface forms), but the second method is language dependent as it requires morphological analyzers, morphological synthesizers and manually created term morpho-syntactic synthesis rules that specify how a term (of a specific morpho-syntactic structure) in a given language has to be inflected in order to acquire all possible surface forms.

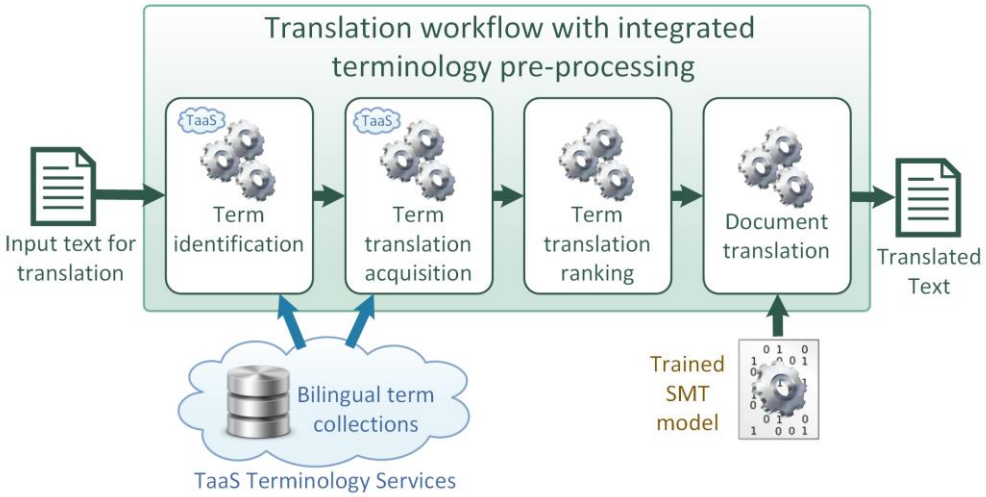

Figure 2. Translatable content pre-processing workflow for dynamic terminology integration in SMT systems

To evaluate the dynamic terminology integration workflow, we have performed automatic evaluation (for four language pairs) using standard SMT evaluation metrics and manual

[3] More details on the *Moses* XML mark-up can be found on the *Moses* SMT system's home page at: http://www.statmt.org/moses/?n=Moses.AdvancedFeatures#ntoc7.

comparative evaluation (for three language pairs) between baseline systems and systems with dynamic terminology integration support. The baseline systems for all languages were trained on the LetsMT platform using the DGT-TM parallel corpus (the releases of 2007, 2011 and 2012) (Steinberger et al., 2012). For evaluation purposes, professional translators were asked to prepare professional term collections in the narrow domain of car service manuals. Then, professional translators were asked to prepare a term collection for all four language pairs (the four different bilingual term collections consist of an average of 654 term pairs) and to translate an evaluation data set from English into the respective languages. The evaluation data set consists of 872 sentences from car service manuals. For English-Latvian, the baseline system was tuned using in-domain tuning data. However, because such an in-domain tuning corpus was not available for the remaining languages, the SMT systems were tuned using 2000 held-out sentence pairs randomly extracted from the training data.

The automatic evaluation results using BLEU and TER metrics are given in Table 2. The results show that the dynamic terminology integration method allows significantly improving baseline SMT system translation quality. The results also suggest that the narrow automotive domain texts contain a significantly different language (in terms of style, word preference, etc.) than found in the SMT system training data. Therefore, the baseline systems achieve a very low translation quality on the car service manual evaluation data. Nevertheless, the dynamic terminology integration, allows translating identified terms better.

Evaluation scenario	English-Latvian		English-Lithuanian		English-Estonian		English-German	
	BLEU	**TER**	**BLEU**	**TER**	**BLEU**	**TER**	**BLEU**	**TER**
Baseline on SMT system's in-domain data	*-*	*-*	*48.12*	*52.02*	*47.81*	*-*	*54.03*	*47.18*
Baseline on car service manuals	12.68	78.01	6.94	90.68	6.26	93.23	8.27	88.12
With dynamic integration	16.09	73.60	7.99	87.98	6.66	89.41	9.17	82.28
Relative improvement	**26.9%**	5.7%	**15.1%**	3.0%	**6.4%**	4.1%	**10.9%**	6.6%

Table 1. Automatic evaluation results of the dynamic terminology integration in SMT systems

Based on the methodology by Skadiņš et al. (2011), translators (in average 7 per language pair) were asked to evaluate translations from the baseline systems and from the dynamic terminology integration scenario. An average of 576 sentences were evaluated per language by all evaluators. The evaluation results in Table 2 show that the dynamic terminology integration allows producing significantly better results than the baseline system. It has to be noted that indecisive situations in the evaluation methodology penalize both systems.

Language pair	System	Total points
English-Latvian	Baseline	40.14±4.0%
	Improved	59.86±4.0%
English-Lithuanian	Baseline	44.19±4.21%
	Improved	55.81±4.21%
English-Estonian	Baseline	45.38±3.93%
	Improved	54.62±3.93%

Table 2. Manual evaluation results of the dynamic terminology integration in SMT systems

8. Application scenarios

The MT system created for the Latvian public sector will have a number of beneficial usage facilities, including:

- public online translation interface

- website translator
- webpage translation widget
- translation API
- mobile translation apps

These various usage facilities will enable a number of important application scenarios.

The online interface is publicly available on the web, hosted in the Latvian public sector's infrastructure. The online translation interface can be used by a variety of audiences: Latvian citizens who wish to translate public sector data, government employees who wish to translate public sector data, and European citizens who want to translate Latvian public sector data into English.

Within this public online interface, users can take advantage of three key functions: text translation, document translation, and website translation. With text translation, users can translate words and sentences. Document translation allows for the translation of entire documents, either by uploading or copying/pasting documents. Translated documents preserve the original formatting, which is particularly beneficial when translation legislative documents with complex subparagraphs.

The website translator is available as an online tool. Using the website translator, users can easily translate the content of any website using. Users simply enter a web address in order to get a translation.

A webpage translation widget is available for integration into public sector websites. The widget allows for light in-place post-editing by web administrators, so the MT system can also be continuously improved and enhanced by administrators. Due to the broad range of domains represented by the system, this is an added bonus, as it allows for specialists in each field to input their corrections to translations. For instance, an administrator at the Ministry of Justice – who possesses special knowledge of the judicial domain – is equipped to correct machine translation of judicial texts, which are henceforth presented as output in the Latvian public sector MT system. Online visitors to the given website can also use the widget to translate content into their own language.

The API allows web administrators to integrate MT into any solution. MT has currently been integrated into the Latvian government's official e-services portal, www.latvija.lv. Thanks to integration into e-government services, both citizens and non-citizens can receive e-services in multiple language – substantially improving and expanding government communication and increasing its reach to a wider array of citizens.

The mobile translation app allows for users to benefit from MT on mobile devices. These include Android, iOS, and Windows Phone. Users can use the mobile app to translate texts and phrases wherever they happen to be – for instance, at government conferences – retrieving translation results from the public sector MT systems.

The MT system build for the Latvian public sector will also play a vital role in the most important event for the Latvian public sector since the country's accession to the EU: Latvia's hosting of the 2015 Presidency of the Council of the European Union. During the six-month presidency, from January to June of 2015, thousands of international journalists and foreign delegates will visit Riga for high-level EU meetings. They will have an acute need for fast, high-quality translation.

For this reason, the public sector's MT system will be made available during the EU Presidency in four facilities:

- Desktop application in the EU Presidency's media center
- MT kiosk, or work station, in the headquarters of the EU Presidency

- Mobile app, available for free download
- Online interface, available on the official website of the EU Presidency

Thanks to these deployments of the Latvian public sector's MT system, the EU Presidency will be able to provide for multilingual communication. This represents the greatest achievement of machine translation – the bridging of language barriers between people and cultures.

The work on the public sector MT systems in the Baltic countries contributes to the broader initiative on the creation of the public Pan-European infrastructure for automated translation services. This infrastructure will provide machine translation services for other digital service infrastructures (DSIs) envisioned by the Connecting Europe Facilities (CEF) Programme. Translation services will help to remove language barriers, facilitate cross-border information exchange and enable cross-border access to online content and services in the EU digital single market.

9. Conclusions

As a result of the project to create MT systems for the public sector, Baltic governments have been enabled with machine translation systems specifically designed to meet public sector needs. This will lower translation costs for the governments by a considerable degree, as well as facilitate communication with all linguistic groups residing in the country. In addition, the system allows for government employees to translate documents into English as well, providing access to information about Latvia and Lithuanian to citizens of other countries.

The MT system have only recently been completed, and work on the projects is still ongoing. Therefore it is too early to identify specific benefits and discuss the outcome of various applications. These still remain to be seen. However, this experience in developing MT systems for Baltic governments can be useful for other countries that are considering building MT systems to enable multilingual communication and meet the needs of their citizens in the digital age.

Acknowledgements

We would like to acknowledge contribution of our colleagues in the work described in this paper – Indra Sāmīte, Valters Šics, Roberts Rozis, Toms Miks, Jānis Teseļskis, Rimas Blažaitis, and Andrius Marcinkevičius. This work has benefitted from the results of the European Union research and innovation projects supported by the ICT PSP programme (LetsMT!) and 7th Framework Programme (MLi, TaaS, ACCURAT).

References

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh

Babych, B., & Hartley, A. (2003). Improving Machine Translation Quality with Automatic Named Entity Recognition. In Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT.

Bojar, O., Mareček, D., Novák, V., Popel, M., Ptáček, J., Rouš, J., & Žabokrtský, Z. (2009). English-Czech MT in 2008. *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athens, Greece: Association for Computational Linguistics

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic.

Carl, M., & Langlais, P. (2002). An Intelligent Terminology Database as a Pre-processor for Statistical Machine Translation. In *COLING-02 on COMPUTERM 2002: second international workshop on computational terminology*-Volume 14 (pp. 1–7).

Daumé, H. (2007). Frustratingly easy domain adaptation. In *the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic.

Foster, G., Goutte, C., and Kuhn, R. (2010). Discriminative instance weighting for domain adaptation in statistical machine translation. In *the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 451–459, Cambridge, MA.

Hálek, O., Rosa, R., Tamchyna, A., & Bojar, O. (2011). Named Entities from Wikipedia for Machine Translation. In *Proceedings of the Conference on Theory and Practice of Information Technologies (ITAT 2011)* (pp. 23–30).

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... & Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (pp. 177-180). Association for Computational Linguistics.

Koehn, P., & Hoang, H. (2007). Factored translation models. In *Proceedings of EMNLP-CoNLL* (pp. 868–876)

Koehn, P., and Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *the Second Workshop on Statistical Machine Translation*, pages 224–227, Stroudsburg, PA, USA.

Koehn, P. (2010). An experimental management system. *The Prague Bulletin of Mathematical Linguistics*, 94.

Matsoukas, S., Rosti, A., and Zhang, B. (2009). Discriminative corpus weight estimation for machine translation. In *the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 708–717, Suntec, Singapore.

Moore, R.C. (2002). Fast and Accurate Sentence Alignment of Bilingual Corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, pp. 135-144. London, UK: Springer-Verlag,

Munteanu, D. and Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Nakov, P. (2008). Improving English-Spanish statistical machine translation: Experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In *the Third Workshop on Statistical Machine Translation*, pages 147–150, Columbus, Ohio.

Papineni, K, Roukos, S, Ward, T, Zhu, W. (2002). BLEU: a method for automatic evaluation of ma-chine translation. *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL)*.

Sennrich, R. (2012). Perplexity minimization for translation model domain adaptation in statistical machine translation. In *the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549, Avignon, France.

Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufis, D., Verlic, M., ... & Pinnis, M. (2012). Collecting and Using Comparable Corpora for Statistical Machine Translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

Skadiņš, R., Puriņš, M., Skadiņa, I., & Vasiļjevs, A. (2011). Evaluation of SMT in Localization to Under-Resourced Inflected Language. In Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT 2011) (pp. 35–40). Leuven, Belgium: European Association for Machine Translation.

Steinberger, R, Eisele, A, Klocek, S, Pilos, S, Schlüter, P. (2012). DGT-TM: A freely Available Translation Memory in 22 Languages. *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012)*, Istanbul.

TaaS. (2014). Public Deliverable D4.4 Integration with SMT Systems. TaaS Project: Terminology as a Service.

TTC. (2013). Public Deliverable D7.3: Evaluation of the Impact of TTC on Statistical MT (p. 38). TTC Project: Terminology Extraction, Translation Tools and Comparable Corpora.

Vasiļjevs, A., Skadiņš, R., & Tiedemann, J. (2012). LetsMT!: Cloud-Based Platform for Do-It-Yourself Machine Translation. In *Proceedings of the ACL 2012 System Demonstrations* (pp. 43–48). Jeju Island, Korea: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P12-3008

Zhao, B., Eck, M., and Vogel, S. (2004). Language model adaptation for statistical machine translation with structured query models. In *the 20th International Conference on Computational Linguistics (Coling 2004)*, page 411, Geneva, Switzerland.