Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/FAIA200621

Language Technology Platform for Public Administration

Raivis SKADIŅŠ [a,b], Mārcis PINNIS [a,b], Artūrs VASIĻEVSKIS [a], Andrejs VASIĻJEVS [a,b,1],
Valters ŠICS [a], Roberts ROZIS [a] and Andis LAGZDIŅŠ [a]
[a] *Tilde, Riga, Latvia*
[b] *Faculty of Computing, University of Latvia, Latvia*

Abstract. The paper describes the Latvian e-government language technology platform HUGO.LV. It provides an instant translation of text snippets, formatting-rich documents and websites, an online computer-assisted translation tool with a built-in translation memory, a website translation widget, speech recognition and speech synthesis services, a terminology management and publishing portal, language data storage, analytics, and data sharing functionality. The paper describes the motivation for the creation of the platform, its main components, architecture, usage statistics, conclusions, and future developments. Evaluation results of language technology tools integrated in the platform are provided.

Keywords. Language technology infrastructure, machine translation, speech recognition, speech synthesis, terminology, language resources

1. Introduction

Machine translation (MT) and other language technologies (LT) are invaluable tools for the public sector to reach out and connect with its various constituents in a cost effective and secure way. Language technologies can simplify, automate, and broaden the way public administration interacts with the public in their language.

Technologies like machine translation can significantly reduce the time and costs of translation [1][2] in public sector institutions. In many scenarios, machine translation is the only feasible way to provide access to e-government services in multiple languages. For instance, MT can be used as an assistive technology for vital information distribution in crisis situations [3].

There is a growing pressure to find an efficient solution to tackle language barriers in the multilingual European Union with its 24 official languages, many of which are spoken by less than 10 million people [4]. This is highlighted in the European Parliament resolution on language equality in the digital age adopted on 11 September 2018 that calls on member states and European Commission to boost the development and application of translation technologies and other LT for all EU languages, including languages that are less widely spoken [5]. The language technology community has proposed development of a Pan-European infrastructure for language tools and services to address the multilingual needs of the public sector, industry and society [6].

[1] Corresponding Author: Andrejs Vasiļjevs; Tilde, Vienibas gatve 75a, Riga, Latvia, LV1004; E-mail: andrejs@tilde.com.

The European Commission, with the support of multiple companies and research organisations, addresses the translation needs of public administrations with its online machine translation service eTranslation. eTranslation provides MT functionality from/to any official EU language. It supports plaintext and formatting-rich document translation in asynchronous translation mode.

The platform approach for addressing multilingual needs in an intergovernmental context is exemplified by the EU Council Presidency Translator[2] – a custom-tailored multi-functional translation solution to support the hosting countries of the presidencies of the Council of the European Union [7]. This machine translation platform supports translation from/to all 24 official European languages. The platform supports plaintext, formatting-rich document, and website translation. Registered users from public administration institutions have access to the SDL Trados Studio plug-in, enabling MT support in CAT tool environments. The initial development of the EU Council Presidency Translator was funded by the European Commission through the Connecting Europe Facility (CEF) Telecom programme.

In Lithuania, a language technology platform *versti.eu* provides similar machine translation functionality to translate plaintext, formatting-rich documents (by supporting the most popular MS Office formats), and website translation. The *versti.eu* platform is freely available without registration and is maintained by Vilnius University.

The government of Latvia is among the pioneers in advancing a platform approach to meet multilingual needs on a national level. For the Latvian government, a particular challenge is to ensure that public information and e-services are accessible to all linguistic groups living in Latvia or having business, cultural, or private relationships within the country. To address the need for an automated solution to the multilingual challenge, a centralized language technology platform has been created. The platform, named HUGO.LV[3], developed by Tilde[4] and maintained by the Culture Information Systems Centre, addresses the multilingual needs of public institutions for their internal and external communication.

The paper further describes the motivation for the creation of the platform HUGO.LV, its main components, architecture, usage statistics, as well as presents conclusions and future developments. Evaluation results of the language technology tools integrated in the platform are provided.

2. Motivation

The goal of the language technology platform is to provide the latest developments in language technology in order to help public administrations:

- To reach various audiences and communities by providing instant access to information and e-government services in various languages;
- To exchange information across borders;
- To provide real-time secure translation of confidential texts, documents and websites;
- To boost operational productivity of translation work in public institutions;

[2] http://presidencymt.eu.

[3] http://www.hugo.lv.

[4] http://www.tilde.com.

- To facilitate access to online information and e-services to disabled people;
- To advance the Latvian language in the digital age by making state-of-the-art language technologies developed in Latvia widely accessible and used.

3. Components of the Platform

The core functionality of the platform is machine translation with various usability and integration tools, automated speech recognition and synthesis, and a terminology management and publishing portal.

3.1. Machine Translation Systems

Neural machine translation (NMT) systems for the HUGO.LV platform were trained iteratively during a timeframe of two years. Therefore, the platform supports multiple NMT decoders, including AmuNMT [8] for models trained using multiplicative long short-term memory (MLSTM) [9] based recurrent neural networks, and Transformer [10] models from Sockeye [11] and Marian [12] toolkits. The platform features a total of 12 NMT systems for translation to/from Latvian, English and Russian in the general domain as well as culture and legal domains. The AmuNMT models were trained using the Nematus [13] toolkit. For training of NMT systems, data were prepared using Tilde's parallel data pre-processing workflows (see [14] for more details). For English-Latvian and Latvian-English NMT general domain systems, morphology driven word splitting [15] was applied instead of the simple byte-pair encoding [16].

The quality of the NMT systems was validated using automatic and manual evaluation methods. For the automatic evaluation, we calculated BLEU [17] scores using the ACCURAT balanced evaluation set[5] [18]. The results for the 12 systems are provided in Table 1. The results show that translation quality according to BLEU is lower when translating into morphologically rich languages. BLEU provides even lower scores when analyzing translations between morphologically rich languages. This can be explained by the fact that both Latvian and Russian allow variations in word order that allow the same sentence to be translated using different syntactic structures. However, error analysis could be performed in future work to assess whether the overall error level is comparable when translating from/to morphologically simpler and more complex languages. The legal and cultural domain systems show subpar translation quality when evaluated on the balanced evaluation set, however, this is expected as these are systems adapted on specific datasets.

We also performed manual comparative evaluation at the time of development of the NMT systems. The evaluation was performed by comparing HUGO.LV general domain systems and Google Translate. The evaluation was carried out on (at that time) current news. The results (see Figure 1) show that the translations of the HUGO.LV general domain NMT systems were more preferred by the evaluators (professional translators) than the translations of Google Translate.

[5] ACCURAT balanced test corpus for under resourced languages, available for download in the META-SHARE repository http://www.meta-share.org.

Table 1. Automatic evaluation results (in terms of BLEU scores) of HUGO.LV NMT systems using the ACCURAT balanced evaluation set

Translation direction	Legal	Culture	General
English-Latvian	25.64	21.46	28.69
Latvian-English	31.87	32.93	34.31
Russian-Latvian	16.71	16.07	16.94
Latvian-Russian	15.02	15.40	15.65

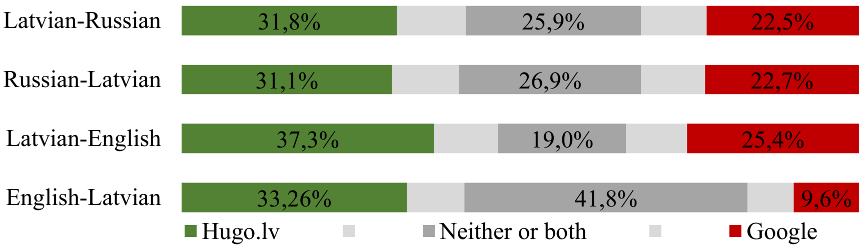

Figure 1. Human comparative evaluation of general domain systems of HUGO.LV and Google Translate

3.2. Text and Document Translation Facilities

The language technology platform provides a translation workspace to translate texts and documents. Users can translate entire documents with a click of a button. Translated documents preserve their original formatting. Multiple formats are supported – rtf, docx, xlsx, pptx, odt, odp, ods, html, etc. A specialized workflow for pre-processing, translating, and post-processing of format-rich documents was developed [19].

3.3. Tools for Website Translation

Two options for website translation are available. A browser add-on for Chrome lets end users to translate any website. For website owners and developers, a translation widget integrates the machine translation functionality into their websites. This lets public administration bodies to provide instant translations of all of their content.

3.4. Tools for Translators

An online computer assisted translation (CAT) tool is integrated in the platform to support semi-professional translation work done by public sector employees. It has been developed by adapting the open-source MateCAT tool [20]. All the segments translated by human translators are stored in a centralized translation memory within the platform. For professional translators, the platform has a plug-in for integrating HUGO.LV machine translation systems in SDL Trados Studio – a computer aided translation tool used by Latvian public administrations.

3.5. Speech Technologies

The text-to-speech functionality provides information for visually impaired or dyslectic people by reading out the written text. Man, woman and youngster voices are provided based on the concatenation approach using diphone synthesis, multiple diphone variations, and LPC residual modification [21].

Automatic speech recognition for Latvian enables text dictation and transcription of audio recordings. It is created with the Kaldi toolkit [22] using an HMM-DNN acoustic model [23], [24] and the Latvian Speech Recognition Corpus [25], [26]. The quality of the Latvian ASR reaches a word error rate (WER) of 9 % as measured on a test corpus.

3.6. Terminology Portal

The terminology portal is a separate platform component that provides an open access to consolidated national terminology resources and supports correct and consistent use of terminology in human and machine translations.

The terminology component has facilities for storing, managing, and accessing national terminology data – 435,000 Latvian terms and 250,000 English terms as well as terms in other languages. Term collections are organized in 22 domains specified by the State Language Center of Latvia. Currently 95 public term collections are available ranging from data digitalized from paper format books and dictionaries to live term collections that are frequently updated by domain and language experts.

The terminology component was created with the following functionality: 1) terminology metadata and term data management; 2) terminology creation workflow; 3) user and their different rights management to ensure online and easy terminology sharing; 4) publishing of news on terminology work, latest protocols and official decisions, some theoretical materials and other content.

The main functionality of the portal is term data and metadata management. All terms are organized in collections. A collection contains concepts that can store the term and its related information in multiple languages. Import and export functionality reuses terminology data in different solutions. TBX, CSV, TSV, MS Excel file format support was created, and these exports are powered with a manual mapping functionality between the file data structure and term database structure. Also, the single term collection view is very important as it provides a full list of term entries within the collection with their data editorial function in place. The terminology portal provides term data export in MT compatible formats for immediate use in training and customising of MT systems.

The terminology creation workflow starts with entering a term candidate and other raw data into the system. When the raw terminology data is prepared, the discussion process can be started. The workflows can be public or private. Public terminology creation workflow enables every Latvian citizen to take part in the discussion about new terms. Private workflows let experts cooperate while keeping the discussions and term candidates confidential. Term creation workflow encourage suggestions for new term translation equivalents, comments on existing ones, as well as comments on a whole terminological concept. During the discussion process, everyone can vote for the best term candidate translation. Finally, the term workflow manager can manually review the list and approve the agreed terms.

Providing content related to the terminology field, the solution helps to form a community of terminologists, and attract their attention with the latest developments in terminology. Also making terminology collections publicly searchable and discoverable allows every citizen to become acquainted with the latest approved terminology.

4. Architecture of the Platform

HUGO.LV is based on the recent version of the LetsMT! platform [27] and has a multi-level architecture (See Figure 2):

- Client level;
- Interface level;
- Logic level;
- Data level;
- High performance computing cluster.

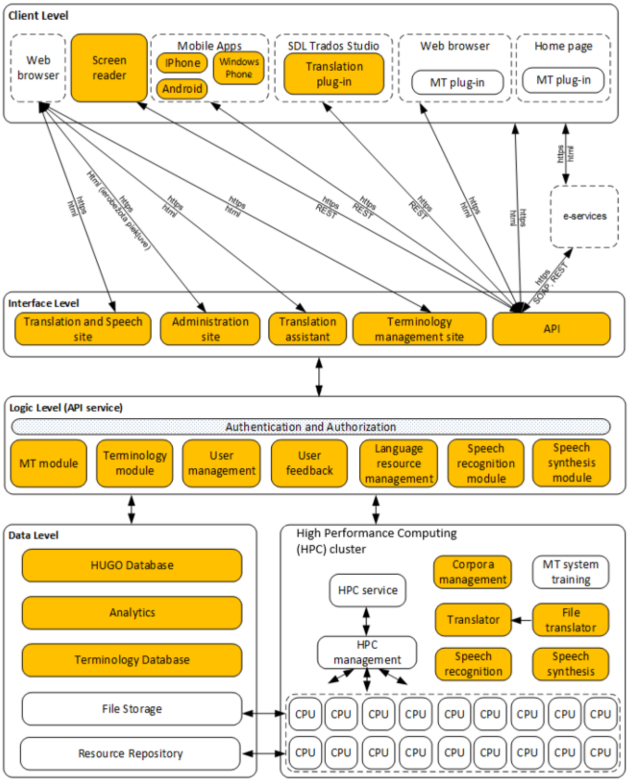

Figure 2. Logical architecture of the platform

4.1. Client Level

The client level includes components that provide HUGO.LV translation functionality on user devices. The client level functionality includes the widget, web browser, and the SDL Trados Studio plug-in components, as well as mobile applications that run on the user's computer or mobile device.

4.2. Interface Level

The interface level includes all system components that are necessary for the system interaction with both human and machine users (the website user interface and the APIs that provide integration across different internal and external systems).

This level provides an interface between external systems, the widget, the web browser plug-in, the translation assistant, and mobile applications. The API, including its OData Service, can be used in external systems. The components at the interface level cooperate with the logic level. The system API has been implemented as a SOAP[6] and/or REST[7] web service (both XML and JSON[8] format). To ensure the security of the data to be transferred, HUGO.LV communication takes place using the HTTPS[9] protocol. The modules are developed in the ASP.NET environment at the interface level.

4.3. Logic Level

The logic level includes modules that provide all functions needed to operate the system. Logic level modules are called only from interface level modules or the 4.5. High Performance Computing (HPC) cluster. Modules are developed in ASP.NET environment that creates separate web services or, in some cases, are included as modules in an interface level application. External users are not allowed to direct access to the level.

4.4. Data Level

The LetsMT! resource repository is used for the storage of language data, their metadata and MT systems. This repository is dedicated to the storage, processing and management of MT language assets. Meanwhile, trained MT systems, which are multiple binary files that together can take several gigabytes, are stored in file storage. Various data that are not directly related to MT systems are stored in the SQL database (MySQL[10]), such as user data, user feedback, terminology, recommended translation fixes, system settings, analytics, etc. External users are not allowed to access this level.

4.5. High Performance Computing Cluster

In order to train MT systems, several model calculations and optimization tasks occur in parallel, which can take from several hours up to two weeks to complete. These computing tasks are performed in a high-performance computing cluster that works on the Oracle Grid Engine[11] platform on the Linux operating system. The HPC cluster performs a variety of processes that require high computing capacities, such as data preparation and processing tasks, text alignment tasks, training tasks for MT systems, translation tasks for texts and files, speech recognition and synthesis tasks. The use of the HPC cluster ensures the scalability of the system, i.e. increasing the performance of the system, if necessary, by automatically adding new computing resources to the HPC cluster during operation.

[6] SOAP: http://www.w3.org/TR/soap/, http://en.wikipedia.org/wiki/SOAP.
[7] REST: http://en.wikipedia.org/wiki/Representational_State_Transfer.
[8] JSON: http://www.json.org/, http://en.wikipedia.org/wiki/JSON.
[9] HTTPS: http://en.wikipedia.org/wiki/HTTP_Secure.
[10] MySQL: http://www.mysql.com/.
[11] Oracle Grid Engine, formerly Sun Grid Engine (SGE): http://gridengine.org/, http://en.wikipedia.org/wiki/Sun_Grid_Engine.

5. Usage of the Platform

Since January 2019, when the fully functional platform was launched, the HUGO.LV website has been visited 1.3 million times, 26.94 million translation requests have been made, and more than 552 million words have been translated. The most frequently used translation direction is from English to Latvian, the second most popular translation direction is from Russian to Latvian.

Speech technologies of the platform have also been popular by users, with 37.70 million words recognized by transcribing 6,686 hours of audio recordings, and 6.2 million words generated using the speech synthesis functionality.

The HUGO.LV machine translation service is integrated in several Latvian government websites, providing multilingual access for information and e-services. Machine translation services are integrated in the state service portal latvija.lv providing descriptions of services in English and Russian languages, Latvian Electronical declaration system eds.vid.gov.lv, electronical auction website izsoles.ta.gov.lv, and the website of the city library of Valmiera biblioteka.valmiera.lv. Speech recognition technology is used by the national radio Latvijas Radio, enabling transcriptions of audio broadcasts in textual form.

The usability of the platform has been recognised in several national and international contests and events. In 2019, the HUGO.LV platform was awarded the “Platinum Mouse”, which is the main award of the IT industry in Latvia, curated by the Latvian Information and Communication Technology Association (LIKTA). In 2015, the HUGO.LV machine translation service was nominated for the World Summit on Information Society (WSIS) Project Prize in the category “Cultural Diversity and Identity, Linguistic Diversity and Local Content”.

6. Conclusions and Future Work

The considerable use of the HUGO.LV services and their integration in various public online systems clearly demonstrate the value and importance of the platform for public administration and society as a whole.

Future activities include expanding the platform with new components for creating multilingual chatbots that can serve multiple public institutions. The chatbots will use new components such as natural language understanding, natural language generation, intent detection, and dialog management, as well as existing components of the platform for machine translation and speech processing.

The technological architecture and modularity of HUGO.LV platform makes it adaptable to other languages and usage contexts. This makes it possible to introduce a similar solution in other countries by using the same framework and integrating the necessary language tools and services. This can significantly boost speed and decrease costs of adapting feature-rich multilingual platform solution across EU member states.

References

[1] Samuel L, Amrhein C, Düggelin P, Gonzalez B, Zwahlen A, Volk M. Post-editing Productivity with Neural Machine Translation: An Empirical Assessment of Speed and Quality in the Banking and Finance Domain. arXiv preprint arXiv:1906.01685. 2019.

[2] Screen B. Productivity and quality when editing machine translation and translation memory outputs: an empirical analysis of English to Welsh translation. Studia Celtica Posnaniensia 2, no. 1; 2017. p. 113-36.

[3] Lewis W, Munro R, Vogel S. Crisis MT: Developing a Cookbook for MT in Crisis Situations. In Proceedings of the Sixth Workshop on Statistical Machine Translation; 2011. p. 501-511.

[4] Rehm G, Uszkoreit H, Dagan I, Goetcherian V, Dogan MU, Váradi T. An update and extension of the META-NET Study "Europe's Languages in the digital age". 2014. p. 1-8.

[5] European Parliament. Resolution of 11 September 2018 on language equality in the digital age (2018/2028(INI)). P8_TA(2018)0332. 2018.

[6] Vasiljevs A, Hajic J, Hummel J, van Genabith J, Kalnins R. European Platform for the Multilingual Digital Single Market: Conceptual Proposal. In Baltic HLT, p. 20-27. 2016.

[7] Pinnis M, Kalniņš R. Developing a neural machine translation service for the 2017-2018 european union presidency. In Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Papers); 2018. p. 72-83.

[8] Junczys-Dowmunt M, Dwojak T, Hoang H. Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. Arxiv. http://arxiv.org/abs/1610.01108. 2016

[9] Krause B, Lu L, Murray I, Renals S. Multiplicative LSTM for sequence modelling. arXiv preprint arXiv:1609.07959. 2016.

[10] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. Advances in Neural Information Processing Systems; 2017. p. 5998-6008.

[11] Hieber F, Domhan T, Denkowski M, Vilar D, Sokolov A, Clifton A, et al. Sockeye: A toolkit for neural machine translation. ArXiv e-prints; 2017.

[12] Junczys-Dowmunt M, Grundkiewicz R, Dwojak T, Hoang H, Heafield K, Neckermann T, et al. Marian: Fast Neural Machine Translation in C++. ArXiv Preprint ArXiv:1804.00344. https://arxiv.org/abs/1804.00344; 2018.

[13] Sennrich R, First O, Cho K, Birch A, Haddow B, Hitschler J, et al. Nematus: a Toolkit for Neural Machine Translation. In Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics; 2017. p. 65-68.

[14] Pinnis M, Rikters M, and Krišlauks R. Tilde's Machine Translation Systems for WMT 2018. Proceedings of the Third Conference on Machine Translation; 2018. p. 477-485.

[15] Pinnis M, Krišlauks R, Deksne D, Miks T. Neural Machine Translation for Morphologically Rich Languages with Improved Sub-word Units and Synthetic Data. Proceedings of TSD 2017: Text, Speech and Dialogue; 2017. p. 237-245

[16] Gage P. A New Algorithm for Data Compression. C Users Journal, 12(2); 1994. p. 23-38.

[17] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a Method for Automatic Evaluation of MT. Proc. of the 40th Annual Meeting on Association for Computational Linguistics; 2002. p. 311-318.

[18] Skadiņš R, Goba K, Šics V. Improving SMT for Baltic Languages with Factored Models. In Frontiers in Artificial Intelligence and Applications, volume 219; 2010. p. 125-132.

[19] Pinnis M, Skadiņš R, Šics V, Miks T. Integration of Neural Machine Translation Systems for Formatting-Rich Document Translation. Natural Language Processing and Information Systems. NLDB 2018. Lecture Notes in Computer Science, vol. 10859; 2018. p. 494-497.

[20] Federico M, Bertoldi N, Cettolo M, Negri M, Turchi M, Trombetti M, et al. The Matecat Tool. In COLING (Demos); 2014. p. 129-132.

[21] Goba K, Vasiļjevs A. Development of Text-To-Speech System for Latvian. Proceedings of NODALIDA 2007, Tartu, Estonia.

[22] Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, et al. The Kaldi speech recognition toolkit. In IEEE 2011 workshop on automatic speech recognition and understanding (No. CONF). IEEE Signal Processing Society; 2011.

[23] Salimbajevs A, Strigins J. Latvian Speech-To-Text Transcription Service. Proceedings of Interspeech 2015; p. 22-723.

[24] Salimbajevs A. Towards the First Dictation System for Latvian Language. Frontiers in Artificial Intelligence and Applications, vol. 289: HLT – The Baltic Perspective; 2016. p. 66-73.

[25] Pinnis M, Auziņa I, Goba K. Designing the Latvian Speech Recognition Corpus. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14); 2014. p. 1547-1553.

[26] Pinnis M, Salimbajevs A, Auzina I. Designing a Speech Corpus for the Development and Evaluation of Dictation Systems in Latvian. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016); 2016. p. 775-780.

[27] Vasiļjevs A, Skadiņš R, Tiedemann J. LetsMT!: A Cloud-Based Platform for Do-It-Yourself Machine Translation. Proceedings of the ACL 2012 System Demonstrations; 2012. p. 43-48.