Developing a Neural Machine Translation Service for the 2017-2018 European Union Presidency

Mārcis Pinnis marcis.pinnis@tilde.com
Rihards Kalnins rihards.kalnins@tilde.com
Tilde, Vienības gatve 75A, Riga, Latvia, LV-1004

Abstract

The paper describes Tilde's work on developing a neural machine translation (NMT) tool for the 2017-2018 Presidency of the Council of the European Union. The tool was developed by combining the European Commission's eTranslation service with a set of customized, domain-adapted NMT systems built by Tilde. The central aim of the tool is to assist staff members, translators, EU delegates, journalists, and other visitors at EU Council Presidency events in Estonia, Bulgaria, and Austria. The paper provides details on the workflow used to collect, filter, clean, normalize, and pre-process data for the NMT systems; and the methods applied for training and adaptation of the NMT systems for the EU Council Presidency. The paper also compares the trained NMT systems to other publicly available MT systems for Estonian and Bulgarian, showing that the custom systems achieve better results than competing systems.

1 Introduction

The administrative work of the European Union (EU) is led by the Presidency of the Council of the EU, which is hosted by a different EU Member State every six months. During its half-year term, the hosting country is tasked with organizing hundreds of high-level events, including conferences and administrative meetings. As the EU Council Presidency brings together delegates and journalists from 28 EU Member States – home to the EU's 24 official languages – the issue of language barriers becomes a major challenge for the politically important event.

To overcome language barriers during the EU Council Presidency in 2017-2018, the Northern Europe-based language technology company Tilde developed a multilingual communication tool that enables automated translation at scale by combining customized neural machine translation (NMT) systems and the Connecting Europe Facility (CEF) eTranslation service for the EU's official languages, developed by the European Commission's Directorate-General for Translation.[1] Tilde first developed a prototype version of the EU Council Presidency Translator for the 2015 EU Council Presidency in Latvia.

eTranslation is a building block of the European Commission's CEF program, which "supports trans-European networks in the sectors of transport, telecommunications and energy"[2] with building blocks that "facilitate the delivery of digital public services across borders."[3] According to the European Commission's CEF Digital website[4],

"the central aim of [the CEF eTranslation service] is to help European and national

[1] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation
[2] https://ec.europa.eu/digital-single-market/en/connecting-europe-facility
[3] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/About+CEF+building+blocks
[4] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/What+is+eTranslation+-+MT@EC+and+eTranslation

> public administrations exchange information across language barriers in the EU, by providing machine translation capabilities that will enable all Digital Service Infrastructures to be multilingual. CEF eTranslation builds on the existing machine translation service of the European Commission, MT@EC, developed by the Directorate-General for Translation (DGT). MT@EC translation engines are trained using the vast Euramis translation memories, comprising over 1 billion sentences in the 24 official EU languages, produced by the translators of the EU institutions over the past decades."

By combining the CEF eTranslation service with custom NMT engines developed by Tilde, the EU Council Presidency Translator is used to translate text snippets, documents, and websites using a responsive online translation website and a computer-assisted translation (CAT) tool plugin. The main users for the translation tools include EU Council Presidency staff members, public sector translators in the hosting country of the Presidency, EU delegates, and international journalists covering the events. The service was first utilized during the 2017 EU Council Presidency in Estonia[5], from July-December 2017, and featured NMT systems for Estonian, a highly inflected, agglutinative language with just 1.5 million native speakers.

The customized NMT systems for Estonian were built utilizing Tilde's methods for developing state-of-the-art NMT systems for complex languages. The methods include extensive data collection, corpus filtering (i.e., noise removal), data pre-processing, unknown phenomena modelling, and training of NMT models with state-of-the-art recurrent neural network architectures. Additionally to the custom NMT systems, the EU Council Presidency Translator provides access to all of the machine translation (MT) systems from the CEF eTranslation service, for translation between the 24 official languages of the EU and English.

In 2018, the EU Council Presidency Translator has been expanded to feature customized NMT systems for Bulgarian[6], to support the Bulgarian EU Council Presidency in January-June of 2018. The tool has been integrated directly into the official website of the Bulgarian EU Council Presidency, eu2018bg.bg, where the site's many users from throughout the world can find the translation tool in the main menu under the heading "Media." The tool will be further expanded and adapted to include customized NMT systems for German to support the upcoming Austrian EU Council Presidency, in the second half of 2018.

To date, the EU Council Presidency Translator has been used to translate over 4.5 million words. This encompasses translation requests made in the last three months of the Estonian EU Council Presidency (following the launch of the tool in late September 2017) and the first two months of the Bulgarian EU Council Presidency (from January 1 to February 19, 2018). 95% of translation requests were made for the customized NMT systems developed by Tilde for Estonian and Bulgarian.

Since its launch in September 2017, the EU Council Presidency Translator has received accolades from the European Commission[7], from staff translators at the EC's translation directorate, from the Ministry for the 2018 Bulgarian EU Council Presidency[8], as well as from the prime ministers of Italy and Greece[9], who were introduced to the tool at the EU Digital Summit in Tallinn, Estonia.

In this paper, we describe the unique multilingual challenges faced by the EU Council Presidency, the Presidency's stated requirements for a multilingual communication tool, and

[5] https://www.translate2017.eu

[6] https://eu2018bg.bg/en/translation

[7] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/2018/01/26/Official+Website+of+2018+EU+Council+Presidency+Integrates+CEF+eTranslation

[8] https://eu2018bg.bg/en/news/354

[9] https://tilde.com/news/eu-council-presidency-begins-using-ai-powered-translation-tool

Tilde's methods for developing NMT engines for complex languages.

The paper is further structured as follows: Section 2 describes the challenges and requirements for the EU Council Presidency Translator; Section 3 provides a general overview of the EU Council Presidency Translator; Section 4 describes the interfaces that can be used by users to access the MT systems of the EU Council Presidencies; Section 5 describes the methods used to develop the custom NMT systems; and Section 6 concludes the paper.

2 Challenges and Requirements for the EU Council Presidency Translator

The numerous EU delegates, international journalists, and foreign visitors at the events organized by each hosting country of the EU Council Presidency represent speakers of (at least the) 24 official languages of the EU. Preparation of documents, press releases, event information, cultural programmes, and other texts in all 24 languages would be a costly endeavour for the hosting country. Therefore, the EU Council Presidency requires machine translation systems that can allow the participating parties to consume all the information produced by the hosting country in its own official language.

In addition, the hosting country gathers these thousands of visitors in its own capital city, where local news and information is produced in the language of the hosting country. Therefore, the EU Council Presidency also required machine translation systems that produced highly fluent translations for the official language of the hosting country (Estonian in 2017, Bulgarian and German in 2018).

To make these systems available for the wide variety of individuals attending official EU Council Presidency events the EU Council Presidency also required the above-mentioned machine translation systems to be made as easily usable as possible, i.e., integrated into user-friendly online tools, including the official website of the EU Council Presidency. These tools should allow for the translation of various types of content: text snippets, full documents (e.g., various OpenDocument[10] or Office Open XML[11] formats), websites, and professional translation files (e.g., Translation Memory eXchange (TMX)[12], XML Localisation Interchange File Format (XLIFF)[13], etc.).

The main challenges posed by these requirements were as follows:

- Integration of CEF eTranslation service for all 24 official EU languages.
- Development of customized NMT systems for the official languages of the hosting countries in 2017-2018 (Estonian, Bulgarian, German).
- Development of user-friendly tools for utilizing the machine translation systems (responsive online interface, integration in the official website of the EU Council Presidency, etc.).
- Development of text, document, website translation functionality.
- Development of a MT plugin for staff and public sector translators in the hosting country to utilize in CAT tools.

3 Infrastructure for the EU Council Presidency Translator

To facilitate translation needs of the EU Council Presidencies, the EU Council Presidency Translator has been developed as a toolkit (see Figure 1 for an overview) that utilizes services

[10] http://opendocumentformat.org/

[11] See ISO/IEC 29500 at http://standards.iso.org/ittf/PubliclyAvailableStandards

[12] http://www.ttt.org/oscarstandards/tmx/tmx13.htm

[13] http://docs.oasis-open.org/xliff

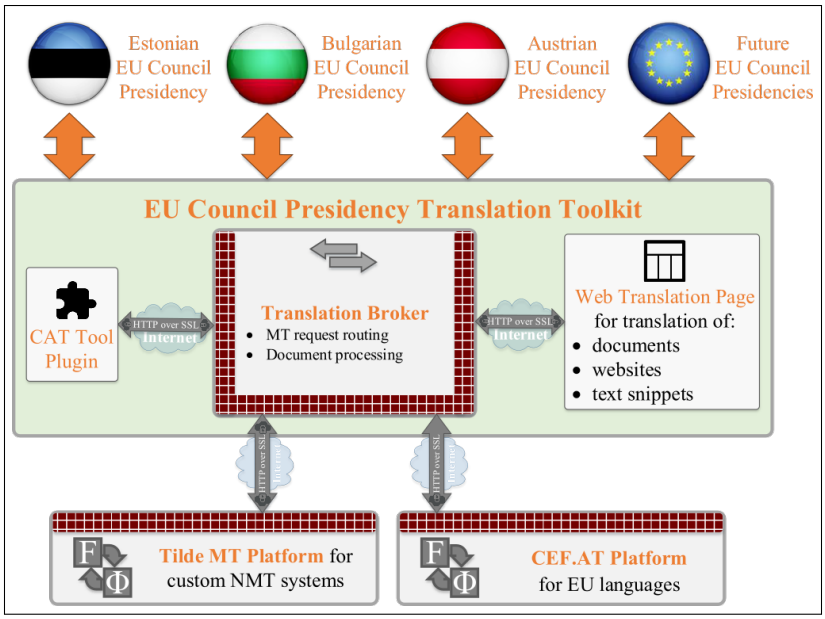

Figure 1: Architecture of the EU Council Presidency Translator

of both the Tilde MT platform (Vasiļjevs et al., 2012) and the CEF eTranslation service. The Tilde MT platform performs two tasks:

1. Serves as an MT system broker that receives translation requests (text snippets, translation segments, documents, and websites) from the two main translation interfaces (the translation website and the CAT tool plugin) and routes the requests to specific MT systems for translation. This architecture allows not only to utilise Tilde MT and eTranslation systems in one user interface, but it also allows to integrate other external MT provider systems within the MT broker in a way that no changes have to be made to the EU Council Presidency Translator's user interfaces.

2. Provides access to the customized EU Council Presidency NMT systems that have been trained and adapted to better translate texts specific to the topics covered by each of the EU Council Presidencies.

4 Translation Interfaces for the EU Council Presidencies

Translators of the EU Council Presidencies, as well as journalists, EU delegates, and other visitors, can use two types of MT interfaces to translate texts, documents, and websites written in the local language of the hosting country into English or content written in English into their own language.

4.1 EU Council Presidency Translation Website

The EU Council Presidency Translator is available online, in a special website linked to the main page of the official EU Council Presidency website.[14] The website is an online translation workspace that allows users to translate texts, full documents (preserving document formatting), and websites in the 24 official EU languages. The portal is customized for each EU Council

[14] https://eu2018bg.bg/en/translation

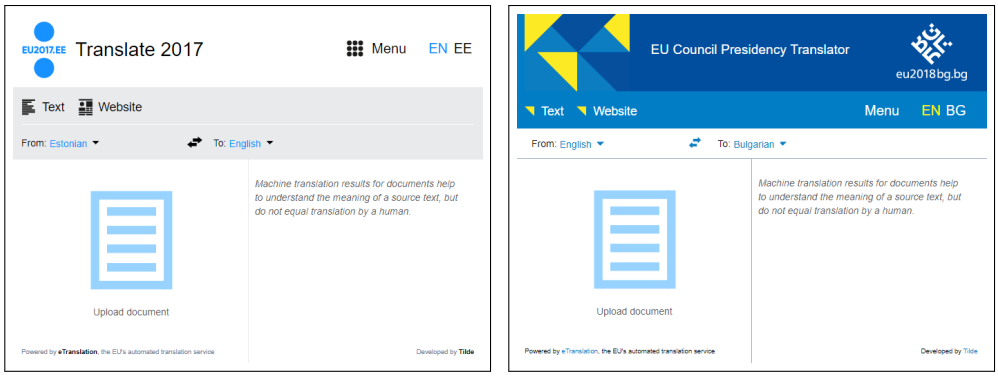

Figure 2: Examples of the Estonian (left) and Bulgarian (right) EU Council Presidency Translator websites

Presidency, featuring the local Presidency branding and other custom elements. An example of the document translation form for both Estonian and Bulgarian EU Council presidencies is depicted in Figure 2.

4.2 CAT Tool Plugin

The EU Council Presidency Translator is also available for professional translators who use the SDL Trados Studio[15] CAT tool with the help of a plugin. The plugin enables users specifically, public administration translators and staff of the EU Council Presidency to utilize the eTranslation service and the custom NMT systems in their everyday work. The plugin provides functionality for translation task pre-translation or translation suggestion preparation in a segment-by-segment translation scenario.

5 Customized NMT Systems

Customized NMT systems for the EU Council Presidencies were developed using the Tilde MT platform, which provides the necessary functionality for corpora cleaning, data pre-processing, and post-processing, as well as allows to deploy NMT systems in a scalable cloud-based infrastructure. The following processing steps were performed to train each of the NMT systems:

- First, parallel and monolingual corpora were collected. The EU Presidency systems have two main goals: 1) to help EU Council Presidency staff members to prepare translations of documents related to the EU Council presidency, and 2) to help visitors of the hosting country to get acquainted with current events taking place in the hosting country. Therefore, the MT systems have two broad target domains - EU Council Presidency and news content. To ensure that the NMT systems are capable of translating such content, focussed web crawling was performed to collect parallel and monolingual data from government and media websites. Additionally, parallel and (in-domain) monolingual data were selected from the Tilde Data Library[16] or supplied by the project's partners.

- Then, the parallel corpora were filtered, cleaned, and normalized using corpora processing tools from Tilde MT. The filtering procedure identifies and removes sentence pairs with the following issues: equal source and target content (i.e., source-source or target-target

[15] http://www.sdl.com

[16] https://tilde.com/products-and-services/machine-translation/features/data-library

entries), sentence splitting issues (e.g., a part of a sentence aligned to a full sentence), foreign (neither source, nor target) language sentences, words, or phrases in source or target sentences, sentence alignment (i.e., non-parallel sentence pair) issues, data redundancy issues, and data corruption issues (e.g., due to errors caused by optical character recognition or wrong formatting of documents). For more details on the filtering procedures, see the paper by Pinnis et al. (2017b). The data were further cleaned and normalized by removing HTML and XML tags, byte order marks, control symbols, escaped characters (e.g., "\n", "\r"), empty braces and curly tags (specific to some CAT tools), decoding XML entities, normalizing whitespace characters and punctuation marks (e.g., quotation marks, apostrophes, dashes, etc.), and separating ligatures (specific to data that are acquired using OCR methods).

- The normalized data were further pre-processed using language-specific tools for non-translatable entity (e.g., e-mail address, file or URL address, various tag and alphanumeric code, etc.) identification, tokenization, and truecasing.
- Following the methodology by Pinnis et al. (2017a), for the Bulgarian EU Presidency we trained NMT models that are more robust to unknown phenomena than vanilla NMT models. To do this, we supplemented the parallel corpus with a synthetic version of the same parallel corpus, which had content words replaced with unknown word tokens in a random manner. To make sure that the same words were replaced on both (source and target) sides, we performed word alignment of the corpus using fast-align (Dyer et al., 2013) and restricted the replacement to only those content words that had non-ambiguous (one-to-one) word alignments.
- Once the data were pre-processed, NMT models were trained using the Nematus (Sennrich et al., 2017) toolkit. All NMT models were sub-word (Sennrich et al., 2015) level attention-based encoder-decoder models with multiplicative long short-term memory units (MLSTM; Krause et al. (2016)). For training, we used the MLSTM model implementation and the NMT training configuration defined by Pinnis et al. (2017b). More specifically, the NMT models were trained using a vocabulary of 25,000 word parts, an embedding layer of 500 dimensions, recurrent layers of 1024 dimensions, dropout rate of 0.2 for recurrent layers and 0.1 for input and output embedding layers, and gradient clipping with a threshold of 1. For parameter updates, the Adadelta (Zeiler, 2012) optimizer with a learning rate of 0.0001 was used.
- When the baseline systems were trained (i.e., they reached the Nematus early stopping criterion of not improving for more than 10 times on validation data), we performed back-translation of in-domain monolingual data in order to prepare synthetic corpora for domain adaptation. The synthetic corpora were filtered, cleaned, normalized and pre-processed using the same workflow that was used to process the parallel corpora. This allowed ensuring that excess noise (i.e., possible NMT mistranslations) was filtered out before performing domain adaptation.
- Finally, NMT model domain adaptation was performed using new training corpora that consisted in balanced proportions (i.e., one-to-one) of the initial training data and the synthetic back-translated data. The one-to-one proportion allows the NMT model to adapt to the required domain, but, at the same time, it allows the model to remember what it had learned during the initial training phase.

Further, we will analyse the data used for training of the NMT systems and the evaluation results of the Estonian (see Section 5.1) and Bulgarian (see Section 5.2) EU Council Presidency NMT systems.

Corpus	English-Estonian	Estonian-English
Cleaned parallel corpus		18,937,780
Cleaned back-translated in-domain corpus	1,716,618	734,417
1-to-1 training data (for domain adaptation)	3,433,236	1,468,834

Table 1: Statistics of the parallel corpora (in terms of unique sentence pairs) used for training of the Estonian EU Council Presidency NMT systems

5.1 NMT Systems for the Estonian EU Council Presidency

The Estonian EU Council Presidency NMT systems were trained on a mixture of publicly available and proprietary corpora. The largest corpora among the publicly available corpora were the Open Subtitles (Tiedemann, 2009) (release of 2016), DTG-TM (Steinberger et al., 2012), Tilde MODEL (Rozis and Skadiņš, 2017), DCEP (Hajlaoui et al., 2014), Microsoft Translation Memories and UI Strings Glossaries (Microsoft, 2015), and Europarl (Koehn, 2005) parallel corpora. The public corpora amounted to approximately half of all the training data. The other half was comprised of proprietary data from the Tilde Data Library.

For domain adaptation, we collected parallel and monolingual corpora from the Estonian EU Council Presidency website[17] and monolingual corpora from various local news agencies. The crawling was restricted to only local resources as the main goals of the EU Council Presidency systems are to enable better translation for content specific to the topics covered by the EU Council Presidency and the topics covered in the news of the hosting country (and not to cater for general translation tasks).

Statistics of the data used for training of the NMT systems are given in Table 1. The data show that for training of the Estonian-English and English-Estonian NMT systems, a substantial amount of data (almost 19 million sentence pairs) were used. The table also shows that the in-domain Estonian monolingual corpus that was used for domain adaptation of the English-Estonian system was more than two times larger than the English monolingual corpus. This is due to the fact that local content in English is much harder to obtain as it is available in much smaller quantities.

The training progress of the baseline NMT systems, as well as the NMT system adaptation process, is depicted in Figure 3. The figure shows that domain adaptation did improve translation quality for the English-Estonian NMT system (by more than one BLEU point), however, the quality increase for the Estonian-English system was rather insignificant (only 0.15 BLEU points). This may be partially explained by the significantly smaller amount of in-domain monolingual data that were available for the creation of the synthetic parallel corpus.

After training and adaptation, we performed automatic evaluation of all NMT systems (both baseline and adapted systems) using BLEU (Papineni et al., 2002), ChrF2 (Popović, 2015), and CharacTER (Wang et al., 2016) (one standard and two newer evaluation metrics that show higher correlation scores to human judgements compared to BLEU for Slavic and Finno-Ugric languages (Bojar et al., 2017)). The results of the evaluation are given in Table 2. The table includes also evaluation results for Google Translate[18] and the English-Estonian and Estonian-English CEF eTranslation systems.

The evaluation was performed using two different evaluation sets: 1) the ACCURAT balanced evaluation set (Skadiņa et al. (2012); a broad domain evaluation set), and 2) an evaluation set created from the parallel corpora of the Estonian EU Council Presidency website (covering also news on various events and topics concerning the Presidency). The results show that both

[17] https://www.eu2017.ee/

[18] https://translate.google.com

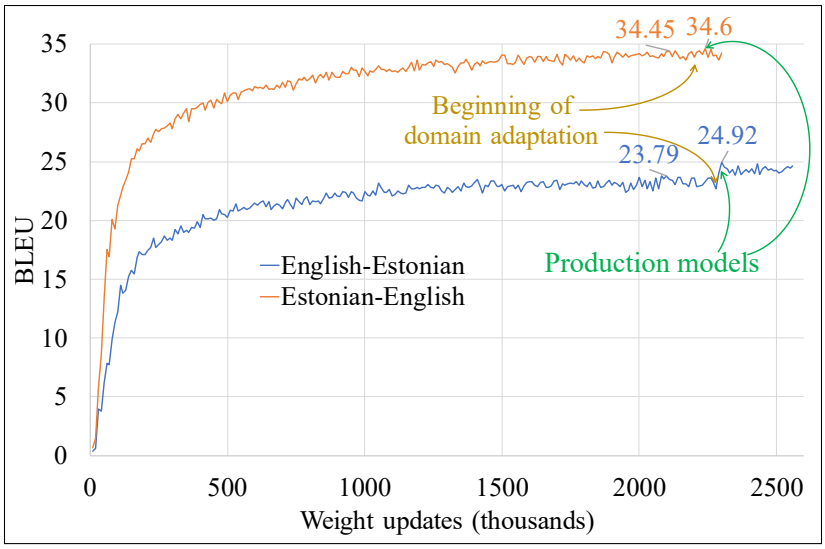

Figure 3: Training progress of the Estonian EU Council Presidency NMT systems

adapted NMT systems achieve the best translation quality on the in-domain evaluation set. I.e., the adapted systems are better suited for translation of texts that cover EU Council Presidency related topics than other compared systems. The results also show that the baseline NMT system for English-Estonian achieves better quality than all competing systems with the adapted system still outperforming other systems also on the broad domain evaluation set. This shows that the adapted NMT systems are also highly competitive broad domain NMT systems.

5.2 NMT Systems for the Bulgarian EU Council Presidency

Similarly to the Estonian NMT systems, the Bulgarian EU Council Presidency NMT systems were also trained on a mixture of publicly available and proprietary corpora. The largest corpora among the publicly available corpora were the DTG-TM, Tilde MODEL, Microsoft Translation Memories and UI Strings Glossaries, DCEP, and Europarl parallel corpora. The public corpora amounted to approximately 79% of all the training data. Slightly over 14% of the data for training of the baseline systems were provided by the project's partners, the Department of Computational Linguistics (DCL) of the Institute for Bulgarian Language (IBL) of the Bulgarian Academy of Sciences.[19] The remaining 7% of the data comprised of proprietary data from the Tilde Data Library. For domain adaptation, monolingual corpora were collected by DCL from various local (for the Bulgarian and English monolingual corpus) and also international (for the English monolingual news corpus) news websites. For the English data, only documents with explicit mentions of Bulgaria were selected.

Statistics of the data used for training of the NMT systems are given in Table 3. The data show that for training of the Bulgarian NMT systems, we used a corpus that was almost 3 times smaller than the parallel corpus that was used to train the baseline NMT systems for the Estonian EU Presidency. However, for the Bulgarian NMT systems, we supplemented the data with synthetically generated data (see Section 5 for details on the synthetic data). Therefore, the total number of sentence pairs that were used for training was almost two times larger than in the initial training data. Because the in-domain monolingual data comprised of approximately the same amount of sentences as the initial training data, the data sets for domain adaptation

[19] http://ibl.bas.bg

System	Broad domain evaluation set			Presidency evaluation set		
	BLEU	ChrF2	Charac-TER	BLEU	ChrF2	Charac-TER
Estonian-English						
Google Translate	**37.85±1.83**	**0.7003**	**0.4592**	31.84±1.26	0.6594	0.5945
eTranslation	37.36±2.76	0.6820	0.4994	28.07±1.21	0.6272	0.6134
PNMT	36.94±1.80	0.6944	0.4604	29.31±1.18	0.6401	0.6014
Adapted PNMT (last model)	35.89±1.84	0.6884	0.4690	**32.46±1.20**	**0.6641**	**0.5410**
Adapted PNMT (best BLEU model)	35.47±1.90	0.6896	0.4665	31.19±1.21	0.6569	0.5592
English-Estonian						
Google Translate	23.23±1.75	0.6553	0.5192	22.72±1.44	0.6409	**0.5063**
eTranslation	24.19±2.15	0.6206	0.5925	20.82±1.40	0.6025	0.5768
PNMT	**25.58±1.66**	**0.6643**	**0.4931**	20.34±1.24	0.6223	0.5420
Adapted PNMT (last model)	24.28±1.64	0.6637	0.4981	**23.18±1.42**	**0.6471**	0.5072
Adapted PNMT (best BLEU model)	23.92±1.60	0.6597	0.5061	22.3±1.29	0.6384	0.5162

Table 2: Automatic evaluation results of the Estonian EU Council Presidency NMT (PNMT) systems

Corpus	English-Bulgarian	Bulgarian-English
Cleaned parallel corpus	6,236,963	
Partner data in the parallel corpus	886,416	
Parallel corpus with synthetic data (for training of the baseline NMT models)	12,116,548	
Cleaned back-translated in-domain corpus	6,188,194	6,098,572
Back-translated corpus with synthetic data	12,068,573	12,209,291
1-to-1 training data (for domain adaptation)	24,325,838	24,185,120

Table 3: Statistics of the parallel corpora (in terms of unique sentence pairs) used for training of the Bulgarian EU Council Presidency NMT systems

reached even 24 million sentence pairs (where approximately 75% amount for all the synthetic data).

The training and adaptation progress for the Bulgarian NMT systems is depicted in Figure 4. The figure shows that (similarly to the trend visible for the Estonian NMT systems) domain adaptation did improve translation quality for the English-Bulgarian NMT system (however, in this case, the improvement was by almost three BLEU points). However, the domain adaptation failed for the Bulgarian-English NMT system. We believe that this may be the result of a too broad coverage of the English monolingual corpus. Because the system has to translate from Bulgarian into English, the English monolingual corpus (for the domain adaptation to work) has to represent what texts in Bulgarian will cover (and not what foreigners may want to write about Bulgaria in English).

The automatic evaluation was performed using two evaluation data sets: 1) a current news evaluation data set, and 2) an EU Council Presidency evaluation data set that covers texts re-

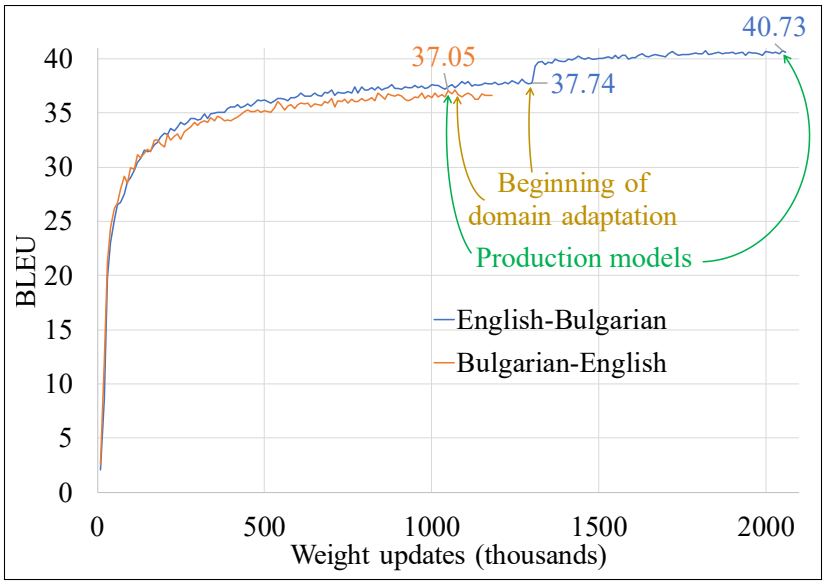

Figure 4: Training progress of the Bulgarian EU Council Presidency NMT systems

System	News evaluation set			Presidency evaluation set		
	BLEU	**ChrF2**	**Charac-TER**	**BLEU**	**ChrF2**	**Charac-TER**
Bulgarian-English						
Google Translate	**38.29±1.15**	**0.6964**	**0.4458**	**46.85±0.91**	**0.7727**	**0.3798**
eTranslation	24.61±0.81	0.6140	0.5725	37.97±0.84	0.7223	0.4565
PNMT	30.84±0.92	0.6579	0.4933	42.68±0.85	0.7486	0.4064
English-Bulgarian						
Google Translate	**38.08±1.32**	**0.6887**	**0.4393**	46.35±0.91	**0.7681**	0.3621
eTranslation	23.36±0.75	0.5977	0.5584	38.03±0.88	0.7230	0.4253
PNMT	31.40±0.96	0.6507	0.4822	44.38±0.85	0.7515	0.3788
Adapted PNMT (best BLEU model)	33.63±1.03	0.6657	0.4659	**46.97±0.90**	0.7672	**0.3620**

Table 4: Automatic evaluation results of the Bulgarian EU Council Presidency NMT (PNMT) systems

lated to the topics covered by the Bulgarian EU Council Presidency. Both evaluation sets were prepared by DLC for evaluation of the EU Council Presidency NMT systems. The results in Table 4 show that for Bulgarian-English the best results are achieved by the Google Translate systems. As mentioned above, domain adaptation for this language pair did not produce better results, which may be the result of domain adherence issues of the monolingual data. However, our baseline NMT systems show significantly better results than the eTranslation systems. This tendency is evident also if we look at the results for the English-Bulgarian systems. However, according to BLEU, the English-Bulgarian adapted NMT system does outperform all other systems on the EU Council Presidency data set. This means that for content covering the EU Council Presidency, the adapted NMT system will be the most suited system.

6 Conclusion

In this paper, we presented the EU Council Presidency Translator developed by Tilde for the 2017-2018 EU Council Presidency. We discussed the architecture of the translation tool and the two main user interfaces - the translation website and the CAT tool plugin. The translation tool is available to all translators, EU delegates, journalists, and other visitors of the EU Council Presidencies in Estonia and Bulgaria. It will also be available for the Austrian EU Council Presidency in the second half of 2018.

In the six months since its launch in September 2017, the EU Council Presidency Translator has helped to translate content amounting to over 4.5 million words (or approximately 470 thousand sentences). The main translation directions for both the Estonian and Bulgarian EU Council Presidencies so far have been between English and the official languages of the hosting countries (amounting to approximately 95% of all translated words).

By applying Tilde's own methods for developing domain-specific NMT systems for complex languages, we were able to create customized NMT systems that outperformed the general eTranslation systems by up to 4 BLEU points for Estonian, and by up to 8 BLEU points for Bulgarian. Tilde's customized NMT systems for Estonian and Bulgarian outperformed Google Translate's general domain NMT engines for the respective language pairs by up to 1 BLEU point.

The tool proves that, when integrated into user-friendly tools, NMT can be successfully applied to enable multilingual communication at high-profile, politically important international events gathering thousands of visitors. The tool also shows that NMT is useful not only for professional translators to boost their productivity, but also as a reading and document analysis tool for a wide range of users in their everyday work, such as EU delegates and international journalists. By applying NMT to their work, users can access information in multiple languages and enjoy better understanding of information, thus helping to promote the aims of goals of high-level events such as the Presidency of the Council of the EU.

7 Acknowledgements

The research has been supported by the European Regional Development Fund within the research project "Neural Network Modelling for Inflected Natural Languages" No. 1.1.1.1/16/A/215.

References

Bojar, O., Graham, Y., and Kamran, A. (2017). Results of the wmt17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, number June, pages 644–648, Atlanta, USA.

Hajlaoui, N., Kolovratnik, D., Väyrynen, J., Steinberger, R., and Varga, D. (2014). Dcep-digital corpus of the european parliament. In *LREC*, pages 3164–3171.

Koehn, P. (2005). Europarl: a parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Krause, B., Lu, L., Murray, I., and Renals, S. (2016). Multiplicative lstm for sequence modelling. *arXiv preprint arXiv:1609.07959*.

Microsoft (2015). Translation and ui strings glossaries.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Pinnis, M., Krišlauks, R., Deksne, D., and Miks, T. (2017a). Neural machine translation for morphologically rich languages with improved sub-word units and synthetic data. In *Proceedings of the 20th International Conference of Text, Speech and Dialogue (TSD2017)*, Prague, Czechia.

Pinnis, M., Krišlauks, R., Miks, T., Deksne, D., and Šics, V. (2017b). Tilde's machine translation systems for wmt 2017. In *Proceedings of the Second Conference on Machine Translation (WMT 2017), Volume 2: Shared Task Papers*, pages 374–381, Copenhagen, Denmark. Association for Computational Linguistics.

Popović, M. (2015). chrf: Character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Rozis, R. and Skadiņš, R. (2017). Tilde model-multilingual open data for eu languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265.

Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Barone, A. V. M., Mokry, J., et al. (2017). Nematus: a toolkit for neural machine translation. *arXiv preprint arXiv:1703.04357*.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2015)*, Berlin, Germany. Association for Computational Linguistics.

Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufi, D., Verlic, M., Vasijevs, A., Babych, B., Clough, P., Gaizauskas, R., Glaros, N., Paramita, M. L., and Pinnis, M. (2012). Collecting and using comparable corpora for statistical machine translation. In Calzolari, N. C. C., Choukri, K., Declerck, T., Doan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 438–445, Istanbul, Turkey. European Language Resources Association (ELRA).

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., and Schlter, P. (2012). Dgt-tm: a freely available translation memory in 22 languages. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 454–459.

Tiedemann, J. (2009). News from opus - a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248.

Vasiļjevs, A., Skadiņš, R., and Tiedemann, J. (2012). LetsMT!: A Cloud-Based Platform for Do-It-Yourself Machine Translation. In Zhang, M., editor, *Proceedings of the ACL 2012 System Demonstrations*, number July, pages 43–48, Jeju Island, Korea. Association for Computational Linguistics.

Wang, W., Peter, J.-t., Rosendahl, H., and Ney, H. (2016). Character : Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation (WMT 2016), Volume 2: Shared Task Papers*, volume 2, pages 505–510, Berlin, Germany.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.