# Tilde MT Platform for Developing Client Specific MT Solutions

**Mārcis Pinnis, Andrejs Vasiļjevs, Rihards Kalniņš, Roberts Rozis, Raivis Skadiņš, Valters Šics**
Tilde
Vienības gatve 75A, Riga, Latvia, LV-1004
{marcis.pinnis, rihards.kalnins, roberts.rozis, raivis.skadins, valters.sics, andrejs}@tilde.lv

**Abstract**

In this paper, we present Tilde MT, a custom machine translation (MT) platform that provides linguistic data storage (parallel, monolingual corpora, multilingual term collections), data cleaning and normalisation, statistical and neural machine translation system training and hosting functionality, as well as wide integration capabilities (a machine user API and popular computer-assisted translation tool plugins). We provide details for the most important features of the platform, as well as elaborate typical MT system training workflows for client-specific MT solution development.

**Keywords:** machine translation, cloud-based platform, data processing workflows

## 1. Introduction

Today, as globalisation and cross-border trade infuse more sectors of the economy, the need for translations and multilingual content is continuously increasing. The demand for translations is surpassing the supply that professional translation services can handle. As Common Sense Advisory reported in a 2016 market study, "enterprises intend to increase their translation volumes by 67% over current levels by 2020." (Lommel and DePalma, 2016) An obvious alternative to customers who cannot afford professional translation services or who have too much content to translate, and also an obvious choice for translation and localisation service providers who see the potential in increasing their productivity, is machine translation (MT). General MT providers (such as Google[1] and Microsoft[2]) cater for the masses with general-domain MT systems. However, customers who require purpose-built and highly customised MT systems, particular for complex or low-resourced languages, turn to MT service providers that offer MT system development and customization capabilities as a full-service package.

One such custom MT platform is Tilde MT[3], the successor of the LetsMT! platform (Vasiļjevs et al., 2012) for Statistical Machine Translation (SMT) system development, first launched in 2012. Tilde MT builds upon the LetsMT! platform by providing greater customisation capabilities, smarter data processing workflows, and current state-of-the-art Neural Machine Translation (NMT) system support.

The paper is further structured as follows: Section 2. provides an overview of the main features of the Tilde MT platform, Section 3. describes the MT system training workflow with the different tools used for data processing, Section 4. describes the MT system translation workflow, and Section 5. concludes the paper.

## 2. Overview of Tilde MT

Tilde MT is a cloud-based custom MT platform that allows users to store linguistic resources (such as parallel and monolingual corpora and multilingual term collections); train SMT and NMT systems; integrate the systems through computer assisted translation (CAT) tool plugins or the Tilde MT external API in users' translation and multilingual content creation workflows; and perform translation of text snippets, documents of various popular formats, and websites directly in the Tilde MT graphical user interface or using a widget that can be integrated in any website.

### 2.1. SMT and NMT System Support

Tilde MT supports two MT paradigms: statistical machine translation and neural machine translation. The platform allows to train Moses phrase-based SMT systems (Koehn et al., 2007) and attention-based encoder-decoder NMT systems with multiplicative long short-term memory units using the Nematus toolkit (Sennrich et al., 2017). The platform has been designed to allow switching to different NMT engines easily. For translation, Nematus NMT models are converted to Marian (formerly AmuNMT) NMT models (Junczys-Dowmunt et al., 2016) that allow reaching much higher translation speed (up to 10 times faster compared to Nematus in non-batched translation scenarios).

### 2.2. Cloud-based Infrastructure

To facilitate on-demand training and deployment of MT systems, it is important for the platform to be highly scalable, available, and reliable. To address these requirements, Tilde MT has been developed as a distributed cloud-based platform (see Figure 1) that is able to dynamically start and turn off computing nodes depending on current workloads. The computing nodes are responsible for running MT system training tasks and translation servers. To provide MT services also to customers with security concerns or customers whose data is not allowed to leave the customers' infrastructure, Tilde MT can be also deployed as an enterprise solution in customer infrastructure.

### 2.3. Tilde Data Library

The platform features a resource-rich data repository, the Tilde Data Library. The library is used as the central data repository for Tilde MT, as well as an open facility for registered users to upload their own corpora (both parallel and

[1] https://translate.google.com
[2] https://www.bing.com/translator
[3] https://tilde.com/mt

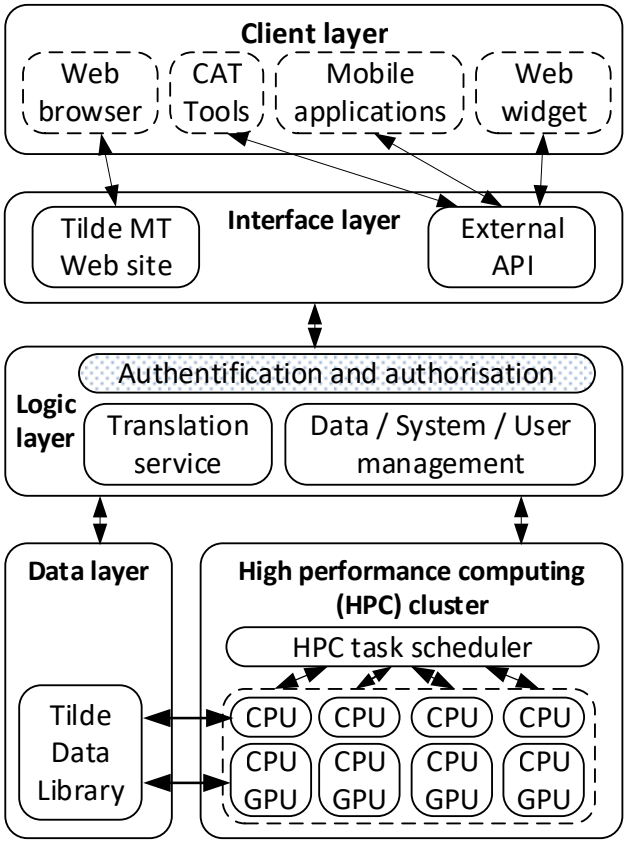


Figure 1: Tilde MT infrastructure design

monolingual) for training MT engines. It stores publicly available and proprietary parallel and monolingual corpora as well as multilingual term collections that users can use to train their MT systems within the Tilde MT platform. Several of the largest (publicly available) parallel corpora in the Tilde Data Library are the DGT-TM (Steinberger et al., 2012), Tilde MODEL (Rozis and Skadiņš, 2017), Open Subtitles (Tiedemann, 2009), MultiUN (Chen and Eisele, 2012), DCEP (Hajlaoui et al., 2014), JRC-Acquis (Steinberger et al., 2006), Europarl (Koehn, 2005), Microsoft Translation Memories and UI Strings Glossaries (Microsoft, 2015). The Tilde Data Library stores approximately 12.35 billion parallel segments for 58 languages and over 4 million terms for more than 125 languages.

### 2.4. Tilde Terminology Integration

Term collections are important linguistic resources that are often used by translators who work on domain-specific translation tasks. To alleviate the need for users to store and manage their terminological resources on multiple platforms, the Tilde MT platform allows users to access their Tilde Terminology (Pinnis et al., 2013) resources. Users can add their term collections to MT system training tasks for static terminology integration as well as to running SMT[4] systems for dynamic terminology integration (Pinnis, 2015). Integration of terminology has been shown to improve term translation accuracy by up to 52.6% (Pinnis, 2015).

[4]For NMT systems, dynamic integration is not yet available.

### 2.5. CAT Tool Plugins and External API

Professional translators often use specific computer-assisted translation tools in their professional duties. Depending on specific projects or customers, translators may have to use different CAT tools. Therefore, it is crucial to the success of an MT platform to provide integration capabilities for at least the most popular CAT tools. MT systems trained on Tilde MT can be accessed from at least four popular CAT tools[5]: MateCat (Federico et al., 2014), SDL Trados Studio, Memsource[6], and memoQ[7].

## 3. MT System Training Workflow

Tilde MT provides users with rich customisation possibilities when training MT systems. Users can specify which filtering and cleaning steps to take, which data pre-processing tools to use, and which MT models and with which configurations to train. Further, we describe in more detail the main MT system training capabilities of Tilde MT.

### 3.1. Data Filtering and Cleaning

Not all data that users upload in the Tilde Data Library as parallel data is actually parallel. For instance, the data may contain misalignment issues, formatting issues, encoding corruption issues, sentence breaking issues, etc. Therefore, the Tilde MT platform performs data filtering before MT system training. The following issues are addressed by various filters in Tilde MT:

1. Source-source or target-target entries in parallel data (equal source/target entries are filtered out).
2. Sentence splitting issues (segments $>$1000 symbols or $>$400 tokens are filtered out; the numerical thresholds here and further can be adjusted for each individual training task).
3. Data corruption through optical character recognition (OCR), e.g., when processing PDF documents (segments containing tokens with $>$50 symbols are filtered out).
4. Redundancy issues (duplicate entries are filtered out).
5. Partial translation (also sentence splitting) issues (entries where the length ratio between the source and target segments is too small (e.g., $<$0.3) are filtered out).
6. Foreign language data issues (entries containing letters from neither source nor target languages are filtered out).
7. Sentence misalignment issues (sentences failing a cross-lingual alignment test using c-eval (Zariņa et al., 2015) are filtered out).

In our previous research, we identified that NMT systems are sensitive to systematic noise that can be found in the parallel data (Pinnis et al., 2017a). Therefore, additionally

[5]For more details, refer to: https://www.tilde.com/products-and-services/machine-translation/features/integration-with-cat

[6]https://memsource.com

[7]https://www.memoq.com

to the filters used for SMT systems, for NMT systems Tilde MT performs also the following filtering steps to ensure that the parallel corpora are filtered more strictly (Pinnis et al., 2017c):

1. Incorrect language filtering using a language detection tool (Shuyo, 2010).
2. Low content overlap filtering using the cross-lingual alignment tool *MPAligner* (Pinnis, 2013).
3. Digit mismatch filtering, which showed to be effective in identifying parallel corpora sentence segmentation issues.

After filtering, each valid segment is cleaned in order to further reduce noise and to remove potential non-translatable text fragments, which will be processed by the formatting tag handling method in the translation workflow prior to decoding but are not necessary during training. The following are the main cleaning steps:

1. Removal of HTML and XML tags,
2. Removal of the byte order mark
3. Removal of escaped characters (e.g., "\n")
4. Decoding of XML entities, normalisation of whitespace characters
5. Removal of empty braces and curly tags (specific to parallel corpora extracted from some CAT tools)
6. Separation of ligatures into letters (specific to parallel corpora extracted using OCR methods)

### 3.2. Data Pre-processing

Next, the filtered and cleaned corpora are pre-processed using standard and custom tools. This step is identical for both the training and translation workflows. The following pre-processing steps are performed:

1. *Normalisation of punctuation*. Tilde MT allows limiting MT models to one standard of quotation marks and apostrophes.
2. *Identification of terminology*. For SMT systems, dynamic terminology integration support ensures that terms can be identified in the source text and possible translation equivalents can be provided to the SMT engine before the actual translation.
3. *Identification of non-translatable entities*. E-mail addresses, URLs, file addresses and XML tags can be identified and replaced with place-holders.
4. *Tokenisation*. Tilde MT uses a regular expression-based tokeniser that allows applying customised tokenisation rules for each language and customer.
5. *Truecasing*. The standard Moses truecasing tool *truecase.perl* can be used to truecase the first or all words of each sentence.
6. *Morphology-driven word splitting* (MWS) (Pinnis et al., 2017b) or *byte-pair encoding* (BPE) (Sennrich et al., 2015). For NMT systems, tokens can be split using a morphological analyser[8] and processed with BPE.
7. *Source side factorisation*. Tilde MT supports NMT models that use linguistic input features (Sennrich and Haddow, 2016). Therefore, the source side can be factored using language-specific factorisation tools (depending on the source language - either part-of-speech or morphological taggers or syntactic parsers).

### 3.3. SMT and NMT Model Training

After the data is pre-processed, SMT or NMT models are trained. During configuration of an MT system, users can freely select whether to train an SMT or an NMT model.
For SMT models, word alignment is performed using fast-align (Dyer et al., 2013), after which a 7-gram translation and the *wbe-msd-bidirectional-fe-allff*[9] reordering models are built. For language modelling, the KenLM (Heafield, 2011) toolkit is used (the n-gram order can be specified by the users). SMT systems are tuned using MERT (Bertoldi et al., 2009).
For NMT models, training data is further pre-processed by introducing unknown phenomena (i.e., unknown word tokens) within training data following the methodology by Pinnis et al. (2017b). Then, an NMT model is trained using the configuration specified by the user (e.g., the vocabulary size, embedding and hidden layer dimensions, whether to use dropout, the learning rate, gradient clipping, etc. parameters can be freely configured). To ensure the stability of the system, a maximum value restriction is applied for each configuration parameter that influences the hardware resources to be consumed.

## 4. Translation Workflow

When an MT model is trained, a translation server can be started. A typical translation server (see Figure 2 for a broad overview) allows to translate text snippets, translation segments (i.e., content that includes tags), documents (e.g., various OpenDocument[10] and Office Open XML[11] formats, translation and localisation formats, such as XML Localisation Interchange File Format (XLIFF)[12] (with different variations), Translation Memory eXchange (TMX)[13], etc.), and web sites (the latter two basically consist of zero to many translation segments).
When translating translation segments, first, formatting tags are removed from the segments and the tag positions are remembered for reinsertion after translation. Then, the text is pre-processed using the same steps that were used for the training data. Additionally to the previous pre-processing steps, the text is also split into sentences (before

[8] Currently, for Latvian and English only.
[9] For more information, see http://www.statmt.org/moses/?n=FactoredTraining.BuildReorderingModel
[10] http://opendocumentformat.org/
[11] See ISO/IEC 29500 at http://standards.iso.org/ittf/PubliclyAvailableStandards
[12] http://docs.oasis-open.org/xliff
[13] http://www.ttt.org/oscarstandards/tmx/tmx13.htm

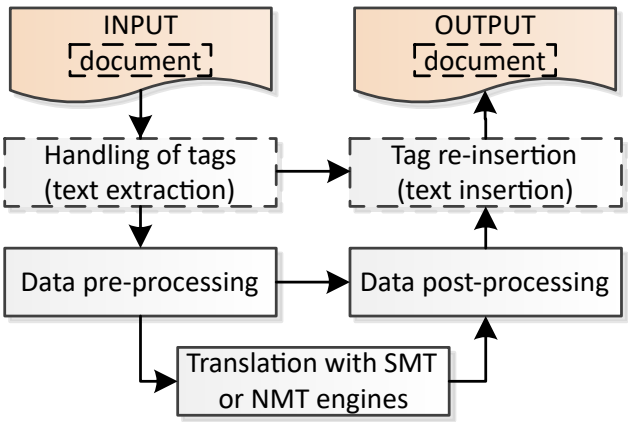


Figure 2: A broad overview of the translation workflow

MWS). For NMT systems, rare and unknown words are identified (based on word part unigram and bigram statistics in the training data) and replaced with unknown word tokens in order to assist the NMT model in the handling of rare and unknown phenomena (Pinnis et al., 2017c). Each pre-processed sentence is then translated using the MT engine that was used during training.

After translation, unknown word tokens are replaced back with their source language tokens, the text is recased and detokenised using language-specific detokenisation rules. Tilde MT provides also functionality for rule-based localisation, i.e., the transformation of certain types of tokens according to customer specified localisation rules. For instance, a customer can pre-set the desired styles of quotation marks and apostrophes, number formats (i.e., decimal point and thousand separators), even conversion rules for different units of measurement (e.g., imperial to metric units, etc.). Finally, if the source was a translation segment, formatting tags are re-inserted in the translated text and (in case of document and Web page translation) the translated segment is inserted in the final document.

In order for the whole translation workflow to work, it is important to keep track of word and phrase alignments and the changes of the word and phrase alignments at each step. When using SMT models, the word alignments are provided by the phrase-based translation model, and when using NMT models, the word alignments are extracted (Pinnis et al., 2017c) from the alignment matrices produced by the attention mechanism of the NMT model.

To address scalability requirements, Tilde MT allows starting multiple translation server instances of each MT system. To save computing resources, translation servers can be set to *fall asleep* after a certain time without any translation requests.

## 5. Conclusion

In this paper, we presented Tilde MT, a distributed cloud-based custom machine translation platform that is capable of supporting SMT and NMT systems. Tilde MT with its feature-rich MT system training workflow alleviates a lot of manual work necessary for data preparation prior to training. The workflows have been specifically adjusted to cater for NMT system, which are more sensitive to systematic noise, development. The platform also allows training more robust NMT models by preparing data in a way that it contains unknown phenomena in common contexts. We also described the translation workflow and its abilities to handle unknown phenomena and to facilitate customer specific customisation needs. Finally, we briefly discussed also the various integration possibilities offered by Tilde MT, such as the CAT tool plugins, the external API, and the Tilde Terminology integration).

## 6. Acknowledgements

In accordance with the contract No. 1.2.1.1/16/A/009 between the "Forest Sector Competence Centre" Ltd. and the Central Finance and Contracting Agency, concluded on 13th of October, 2016, the study is conducted by Tilde Ltd. with support from the European Regional Development Fund (ERDF) within the framework of the project "Forest Sector Competence Centre".

## 7. Bibliographical References

Bertoldi, N., Haddow, B., and Fouet, J.-B. (2009). Improved Minimum Error Rate Training in Moses. *The Prague Bulletin of Mathematical Linguistics*, 91(1):7–16.

Chen, Y. and Eisele, A. (2012). MultiUN v2: UN Documents with Multilingual Alignments. In *LREC*, pages 2500–2504.

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.

Federico, M., Bertoldi, N., Cettolo, M., Negri, M., Turchi, M., Trombetti, M., Cattelan, A., Farina, A., Lupinetti, D., Martines, A., et al. (2014). The MateCat Tool. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 129–132.

Hajlaoui, N., Kolovratnik, D., Väyrynen, J., Steinberger, R., and Varga, D. (2014). DCEP-Digital Corpus of the European Parliament. In *LREC*, pages 3164–3171.

Heafield, K. (2011). KenLM : Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, number 2009, pages 187–197. Association for Computational Linguistics.

Junczys-Dowmunt, M., Dwojak, T., and Hoang, H. (2016). Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. *arXiv preprint arXiv:1610.01108*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and 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*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: a Parallel Corpus for Statistical Machine Translation. In *MT summit*, volume 5, pages 79–86.

Lommel, A. R. and DePalma, D. A. (2016). Europe's Leading Role in Machine Translation. Technical report, Common Sense Advisory.

Microsoft. (2015). Translation and UI Strings Glossaries.

Pinnis, M., Gornostay, T., Skadiņš, R., and Vasiļjevs, A. (2013). Online Platform for Extracting, Managing, and Utilising Multilingual Terminology. In *Proceedings of the Third Biennial Conference on Electronic Lexicography, eLex 2013*, pages 122–131, Tallinn, Estonia. Trojina, Institute for Applied Slovene Studies (Ljubljana, Slovenia) / Eesti Keele Instituut (Tallinn, Estonia).

Pinnis, M., Krišlauks, R., Deksne, D., and Miks, T. (2017a). Evaluation of Neural Machine Translation for Highly Inflected and Small Languages. In *Proceedings of the 18th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2017)*, Budapest, Hungary.

Pinnis, M., Krišlauks, R., Deksne, D., and Miks, T. (2017b). 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. (2017c). 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.

Pinnis, M. (2013). Context Independent Term Mapper for European Languages. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2013)*, pages 562–570, Hissar, Bulgaria.

Pinnis, M. (2015). Dynamic Terminology Integration Methods in Statistical Machine Translation. In *Proceedings of the Eighteenth Annual Conference of the European Association for Machine Translation (EAMT 2015)*, pages 89–96, Antalya, Turkey. European Association for Machine Translation.

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. and Haddow, B. (2016). Linguistic Input Features Improve Neural Machine Translation. In *Proceedings of the First Conference on Machine Translation (WMT 2016) - Volume 1: Research Papers*, pages 83–91.

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.

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*.

Shuyo, N. (2010). Language Detection Library for Java.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The JRC-Acquis: a Multilingual Aligned Parallel Corpus with 20+ Languages. *arXiv preprint cs/0609058*.

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 Min Zhang, editor, *Proceedings of the ACL 2012 System Demonstrations*, number July, pages 43–48, Jeju Island, Korea. Association for Computational Linguistics.

Zariņa, I., Ņikiforovs, P., and Skadiņš, R. (2015). Word alignment based parallel corpora evaluation and cleaning using machine learning techniques. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*.