Customized Neural Machine Translation Systems for the Swiss Legal Domain

Rubén Martínez Domínguez ruben.martinez@tilde.com
Matīss Rikters matiss.rikters@tilde.lv
Artūrs Vasiļevskis arturs.vasilevskis@tilde.com
Mārcis Pinnis marcis.pinnis@tilde.com
Tilde, Vienības gatve 75A, Riga, Latvia, LV-1004

Paula Reichenberg paula.reichenberg@hieronymus.ch
Hieronymus, Stauffacherstrasse 100 CH-8004 Zürich, Switzerland

Abstract

This paper describes Tilde's work on the development of a Neural Machine Translation (NMT) platform for Hieronymus, a Switzerland-based boutique legal and financial translation provider, giving particular attention to the increase in efficiency as regards internal translation processes, as well as NMT's impact on the customer experience of their partners. The NMT tool was developed by combining a set of domain-adapted NMT systems with a customized translation platform, both of which were built and developed by Tilde. The central aim of the solution is to assist Hieronymus translators and to create LexMachina, a secure, do-it-yourself NMT solution for Swiss lawyers. The current paper outlines the workflow used to collect, filter, clean, normalize, and pre-process data for the NMT systems, as well as the methods utilized to train and adapt the NMT systems for Hieronymus. The current paper also sheds light on the needs of Tilde's partner, from approaches to resolving the challenges they faced to the implementation process itself.

1 Introduction

As a steadily growing company in a highly competitive language service industry, Tilde's partner, Hieronymus[1], was eager to adopt an innovative Neural Machine Translation (NMT) strategy to ensure its long-term growth while vastly improving the translation customer experience, as well as gaining the loyalty of their customers via a 'self-service' tool made to cover their needs. Much to their surprise the absence of readily available solutions on the market coupled with the Swiss-specific language context posed considerable challenges when adopting the chosen NMT strategy. Namely, language tools and NMT systems required by Hieronymus had to account for the linguistic specifics of Switzerland's local languages: Swiss-German, Swiss-French and Swiss-Italian. Unsurprisingly, most of the parallel data available for NMT training is in standard German, French, and Italian. Furthermore, Hieronymus' interests lie translation for highly-technical domains: criminal law, tax law, banking, and finance. The NMT systems used by Hieronymus must therefore be able to deliver reliable and trustworthy translations of highly technical domain-specific terminology. Additionally, the NMT systems must be fully integrated into their translation workflows, so as to boost both the internal and external opera-

[1] www.hieronymus.ch

tional efficiency. The NMT integration sought to provide Hieronymus with a competitive edge by streamlining the translation processes while enhancing quality and terminological accuracy.

Another priority of Hieronymus' NMT development was to offer a new product to their clients, mainly law firms, it being a way to enhance the customer experience. Central to this new product was a self-service legal machine translation infrastructure that their partners could independently access with the guarantee of full confidentiality and the "Swiss touch"—two essential elements for Hieronymus' clients.

With the above elements in mind, Tilde combined its language and client-oriented approach with the latest, AI-driven natural language processing technology to develop *LexMachina* [2]. The development of LexMachina was a joint effort between Tilde and Hieronymus, where much attention was placed on the selection and preparation of the right data, as well as the testing and improvement of the same. *LexMachina* is a customized translation platform that guarantees the security and confidentiality throughout the translation process. It has been launched as a collection of 10 customized NMT systems and will be extended to include new domain-specific NMT systems in the near future.

The platform is based on the Tilde MT platform (Pinnis et al., 2018) and LetsMT technology (Vasiļjevs et al., 2012). It supports multiple input formats and maintains tag and formatting integrity when translating documents. Additionally, the translation platform integrates Hieronymus translation memories (TMs), supports integration of NMT systems into the most commonly used computer-assisted translation (CAT) tools, and allows for integration of the NMT engines into Microsoft Outlook. As a result, the *LexMachina* platform allows Swiss lawyers to instantly translate legal documents in the necessary confidential environment while reaping the benefits of customized NMT technologies. The solution developed by Tilde and Hieronymus may also be adapted to the specific needs of Swiss banks, insurance companies and major advisory and accounting companies.

All NMT systems were tailor-made to conform to Hieronymus' requirements regarding Swiss local language and domain-specific terminology. To that end, we set out to acquire, classify, and align Swiss domain data, reviewing the main details and preparing the correct training formula for the customization thereafter. As a result, alongside Hieronymus, we developed generic Swiss legal engines. Further development on this project will see the release of additional Swiss legal engines specialized in various sub-domains (criminal law, financial law, tax law, etc.).

The current paper describes the development of *LexMachina*, and how Hieronymus leveraged their machine translation capability to increase both productivity and efficiency, allowing them to streamline translation processes and become the first provider to offer a do-it-yourself, legal machine translation solution for Swiss lawyers. In presenting this use case, we bring to light the details of the technological, infrastructural, and linguistic challenges we have experienced, and indeed overcome, while creating and implementing this NMT project. The application of the developed NMT systems aim at facilitating the vision of Tilde's partner, and enable the desired innovation with the creation of customized NMT systems and a self-service translation platform.

2 Requirements

Hieronymus' demand for NMT solutions were not satisfied with those currently available on the market. Most available engines are based on standard German, French, and Italian, omitting essential local elements such as punctuation, vocabulary, lexicon, style, register, grammar structure, and terminology. These differences between Swiss local and standard languages were of particular concern to Hieronymus' customers, among which are local law firms, banks, in-

[2] www.lex-machina.ch

surance companies, and other financial institutions, all of which consider the accuracy of terminology essential.

Thus Hieronymus presented Tilde with a list of requirements that the NMT and translation platform had to meet to be considered adapted to their customers' needs. These needs were primarily a question of data; the NMT systems should be built using in-domain terminology, such as legislative acts and laws, and financial and tax content, adapting them to the specificities of the Swiss-German, Swiss-French, and Swiss-Italian languages.

Due to the nature of work of Hieronymus' customers, all information and documents had to be translated securely. Specifically, it was paramount that the MT system guarantee the confidentiality of sensitive data at all times, and that all data be stored within a Swiss infrastructure environment never to be transferred outside of Switzerland. To reinforce the confidentiality of the translation process, the NMT engines and the *LexMachina* translation platform are hosted in a secure, Swiss-based cloud environment controlled by Hieronymus.

To address the above requirements, Tilde and Hieronymus developed *LexMachina*. *LexMachina* is a set of adapted NMT systems which are integrated into a customized translation platform based on Tilde's MT platform. *LexMachina* provides the following functionalities:

- translation of text snippets (words, sentences, up to several paragraphs);
- translation of documents by preserving formatting and document formats;
- translation of websites by preserving website structure and design;
- CAT tool plug-ins for SDL Tradus Studio and Wordbee.

3 Machine Translation Systems

A typical development cycle of domain-specific MT systems involves MT training on general domain data and adaptation on domain-specific data. The Hieronymus case is different, as the final quality and appropriateness of the MT systems depend not only on their ability to translate domain-specific texts, but also on their being tailored to Swiss language specificities. The following section (3.1) describes how we tackled additional challenges posed by data sparsity, which is result of both occupying a niche domain and Swiss language needs.

3.1 Data Collection

To develop NMT systems for the Swiss legal domain, we used three types of data:

- **Publicly available parallel corpora**. Most publicly available parallel data comprise texts in standard French, Italian, and German. These data are not necessarily of Swiss origin and usually do not contain texts of Swiss German, Swiss Italian, and Swiss French. However, such data are available in large proportions and can help to form baseline models. The largest of such is available from the DGT Translation Memories (Steinberger et al., 2012), Digital Corpus of the European Parliament (Hajlaoui et al., 2014), the Tilde MODEL corpus (Rozis and Skadiņš, 2017), Europarl (Koehn, 2005), and other sources available from the Tilde Data Library[3].
- **Parallel data crawled and extracted from legal-domain Web sites** of institutions of Swiss origin. Having four official languages, many Swiss institutions provide multilingual information on their Websites, making it a valuable asset for machine translation. Therefore, we crawled public institution websites using a parallel data crawler, downloaded

[3] https://www.tilde.com/products-and-services/data-library

monolingual documents, and performed cross-lingual alignment with consecutive parallel data extraction to acquire parallel corpora.

- **Translation memories from Tilde's partner**. The in-domain data that were used to fine-tune NMT systems were provided by Hieronymus, thereby ensuring that the trained NMT systems are tailored specifically to the Swiss language context.

3.2 NMT System Training and Domain Adaptation

For the training of NMT models, we use the Marian NMT toolkit (Junczys-Dowmunt et al., 2018) as it provides the most efficient implementation for training and inference of any standard NMT model. We use Marian's standard configuration[4] of the *transformer-base* model (Vaswani et al., 2017). We select training batch sizes dynamically so that they fit in a workspace of 9,000-22,500 MB (depending on GPU specification). We train models with early stopping (Prechelt, 1998), using ten consecutive evaluations with no improvement in translation quality on the development set as the stopping criterion. The high level view of NMT system training:

1. First, pre-process all data using Tilde's parallel data pre-processing pipeline (Pinnis et al., 2018), which involves custom-made processes for parallel data filtering, normalization, non-translatable entity identification, tokenization, and truecasing, as well as standard processes for word splitting, and cross-lingual word alignment.

2. Then, for each domain, we perform careful data selection. We split data into four parts: out-of-domain colloquial data, out-of-domain formal language data, out-of-domain Swiss data, in-domain Swiss data. All Swiss data are up-sampled while the colloquial data are down-sampled or discarded. See Table 1 for the summary of training data size for each language pair.

3. Before training, we separate random subsets of 2000 and 1000 parallel sentences from the in-domain Swiss data to be used as development and evaluation data sets, respectively.

4. To make MT models more robust against incomplete or incorrect input, we synthesize additional training data by randomly replacing 1-3 content words in sentences with a placeholder (Pinnis et al., 2017).

5. We train baseline Transformer NMT models with guided alignment using the Marian NMT toolkit. We provide subword-unit-based statistical alignments as an additional input data stream for learning guided alignments, which are important for formatting-rich document translation and integration in computer-assisted translation tools.

6. Finally, we adapt the systems, thereby ensuring conformity to Swiss language specificities and style. Domain adaptation is performed using a 1-1 mix of in-domain Swiss data with an equal amount randomly sampled from the remaining data.

3.3 NMT System Quality

Figure 1 gives results of automatic evaluation of translation quality of *LexMachina* MT systems using BLEU (Papineni et al., 2002) metric. The performance of publicly available Google Translate general domain systems is given for the reference. Results show that *LexMachina* MT systems yield substantially better quality (12.3 BLEU higher on average) than the publicly available counterparts. The substantial difference in performance suggests that the strategy to approaching Hieronymus' requirements for Swiss language and domain-specific MT systems as a two-fold domain adaptation problem has been successful.

[4] https://github.com/marian-nmt/marian-examples/tree/master/transformer

			Baseline Data		Domain Adaptation Data	
			Parallel	Synthetic	Parallel	Synthetic
FR	↔	**EN**	53.5	50.7	0.24	0.22
DE	↔	**IT**	15.1	13.4	0.17	0.14
IT	↔	**FR**	16.6	16.2	0.17	0.16
FR	↔	**DE**	9.6	7.6	1.8	1.4

Table 1: Training data sizes in millions of sentences.

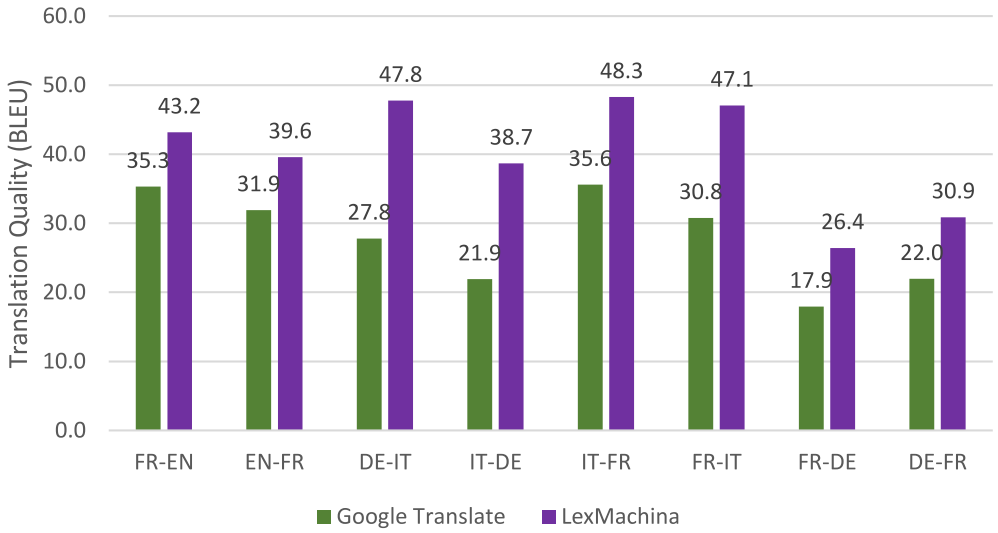

Figure 1: Results of automatic evaluation of translation quality measured in BLEU scores. *LexMachina* in-domain MT systems compared against publicly available general domain MT systems by Google Translate. The comparison was made in February 2020.

4 Implementation

The implementation process for the project was divided in four steps: 1) a pilot to assess NMT capabilities for one language pair, 2) NMT system training, 3) development of the *LexMachina* platform, and 4) deployment of the *LexMachina* platform in Hieronymus' infrastructure. The pilot allowed us to better understand the domain, identify data sources, and establish the domain adaptation strategy for NMT system training. Once satisfied with the results of the pilot, we trained all remaining NMT systems using the strategy established in the pilot phase. The NMT systems were at first deployed on the Tilde MT platform to allow instant access to testing and evaluation of the NMT systems and features of the MT platform. All systems were tested and custom-tweaked by adjusting data pre-processing and post-processing rules. The platform was simultaneously developed according to Hieronymus' requirements. Finally, the platform was deployed in a Switzerland-based, secure data center to comply with the security requirements of Hieronymus and their customers.

The project allowed Hieronymus to reach the following milestones:

- to integrate custom NMT engines in their workflow, which allows their translators to increase productivity and efficiency;

- to become the first provider to offer a self-service, legal machine translation solution for Swiss lawyers;
- to become the first provider to offer a fully secure NMT solution deployed in the Swiss Azure cloud for banks, insurance companies, and major advisory and accounting companies.

5 Conclusions

In response to growing interest from the Swiss banking and insurance industry, both of which want their own specialized NMT engines, Hieronymus and Tilde have developed a common solution to cater for the industry's urgent NMT needs – the current *LexMachina* infrastructure is a proof-of-concept. As a result of the joint project between both parties, Hieronymus can build on the deployed solution and offer on-premises custom NMT engines using both precious corpora developed by Hieronymus, as well as Tilde's extensive experience in setting-up secure infrastructures. Custom-made NMT solutions will allow banks, insurance companies, and large consulting and accounting firms to reduce their translation costs by 30%-50%, improving the quality and speed of delivery - all while maintaining security and confidentiality.

6 Acknowledgements

We would like to acknowledge the contribution, support, and involvement of Hieronymus in the project described in this paper, especially to Orane Laeri and Lauren Spencer, who managed the project on Hieronymus' side. We would also like to thank our colleagues Roberts Rozis, Valters Šics, Igors Zotovs and Viktorija Kononova for their contribution to the project described in this paper.

References

Hajlaoui, N., Kolovratnik, D., Väyrynen, J., Steinberger, R., and Varga, D. (2014). DCEP-Digital Corpus of the European Parliament. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Fikri Aji, A., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Koehn, P. (2005). Europarl : A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit (MT Summit)*, pages 79–86.

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. (2017). 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)*, volume 10415 LNAI, Prague, Czechia.

Pinnis, M., Vasiļjevs, A., Kalniņš, R., Rozis, R., Skadiņš, R., and Šics, V. (2018). Tilde MT Platform for Developing Client Specific MT Solutions. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Prechelt, L. (1998). Early Stopping- but When? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer.

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.

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., and Schlüter, P. (2012). DGT-TM: A Freely Available Translation Memory in 22 Languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 454–459.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention Is All You Need. In *Advances in neural information processing systems*, pages 5998–6008.