How to Move to Neural Machine Translation for Enterprise-Scale Programs—An Early Adoption Case Study

Tanja Schmidt
Welocalize, Inc.
Frederick, MD, United States
tanja.schmidt@welocalize.com

Lena Marg
Welocalize, Inc.
Frederick, MD, United States
lena.marg@welocalize.com

Abstract

While Neural Machine Translation (NMT) technology has been around for a few years now in research and development, it is still in its infancy when it comes to customization readiness and experience with implementation on an enterprise scale with Language Service Providers (LSPs). For large, multi-language LSPs, it is therefore not only important to stay up-to-date on latest research on the technology as such, the best use cases, as well as main advantages and disadvantages. Moreover, due to this infancy, the challenges encountered during an early adoption of the technology in an enterprise-scale translation program are of a very practical and concrete nature and range from the quality of the NMT output over availability of language pairs in (customizable) NMT systems to additional translation workflow investments and considerations with regard to involving the supply chain.
In an attempt to outline the above challenges and possible approaches to overcome them, this paper describes the migration of an established enterprise-scale machine translation program of 28 language pairs with post-editing from a Statistical Machine Translation (SMT) setup to NMT.

1 Introduction

The idea of using recurrent neural networks for machine translation was first presented by Kalchbrenner and Blunsom (2013), followed soon after by Cho et al. (2014) and Sutskever et al. (2014). In a mere three years from those papers, NMT systems were outperforming SMT systems for several translation tasks at the Association for Computational Linguists' Conference on Machine Translation (WMT). At the same time, large translation providers such as Systran, Google and Microsoft announced deployments of NMT systems for public consumption. The combination of these factors quickly made both buyers and providers of translation services aware of the new opportunities.

The rapid emergence of NMT has necessitated that LSPs focus on many new areas, including: qualitative evaluation of individual NMT systems, comparing translation quality and productivity of NMT and SMT systems, implementation and deployment of NMT systems, and building customized NMT systems for specific domains and/or clients.

2 Contextualization

Since the deployment of machine translation technology for commercial use, and especially the breakthrough of Statistical MT solutions, requests for MT as part of regular translation programs have constantly been on the rise. An explosion in the amount of content published as well as increasing pressure to publish content fast and simultaneously in different target markets and languages have caused clients to look into alternative, cheaper options and LSPs to adjust their translation workflows and processes. The continually improving quality of MT systems and new developments such as NMT add to this demand.

As a major global LSP, we count a range of big global companies among our end clients, for whom we typically provide ongoing, on-demand translation services into 20+ languages, covering various content types (= enterprise-scale translation program). It is our role to advise our clients on new developments in (MT) technology, opportunities for automation and

Pérez-Ortiz, Sánchez-Martínez, Esplà-Gomis, Popović, Rico, Martins, Van den Bogaert, Forcada (eds.)
Proceedings of the 21st Annual Conference of the European Association for Machine Translation, p. 309–313
Alacant, Spain, May 2018.

workflow improvement as well as cost and time savings in their translation needs. In our case, we do not work with one specific MT provider, but recommend the MT solution we consider the best fit for a given end client, based on their specific needs and setup.

The arrival of NMT therefore requires us to reevaluate existing MT programs as well as the MT solutions offered by different providers we work with.

3 Planning for NMT for Enterprise-Scale Programs

Since its breakthrough, NMT quickly showed great promise to be able to deliver noticeably higher quality raw machine translations, especially for historically challenging and expensive translation pairs like English-Japanese. In our planning, we therefore started to evaluate a range of the then available, initially generic NMT systems for their qualitative performance on a subset of languages.

The evaluation of these *generic NMT* systems was performed with suitable test content from clients that gave us their permission to use their content for this purpose. We compared these generic systems with the *existing, customized SMT* solutions that were in place for the respective client programs, using automatic scoring for BLEU, GTM, Nist, Meteor, Precision, Recall, TER and Edit Distance (Levenshtein[1]), a post-editing test and human evaluations (see 3.2 Evaluation Methodology for details). While generic NMT frequently outperformed customized SMT on various metrics, the results were inconsistent across content types and languages. Lacking lexical coverage from the generic systems added to this picture, with some languages benefitting more from the increased fluency and grammatical accuracy of the NMT system (e.g. Japanese) while other languages seemed to struggle more with the terminological inaccuracies (e.g. German), at least from a human evaluation viewpoint. Selected results from this study were presented during the 2017 Machine Translation Summit in Nagoya, Japan, the 2017 School of Advanced Technologies for Translators in Trento, Italy, and with the Translation Automation User Society's (TAUS) MT user group (Marg et al., 2017a,b). While results were still mixed at this early stage, they showed that, for some languages, already the generic NMT systems were performing equally well when compared with the established, customized SMT systems. With MT providers starting to make customizable NMT solutions available and the promise in relation to an even better performance from these, we then progressed to direct comparisons on *custom SMT* to *custom NMT*, partly in the form of official client pilot projects.

In the following paragraphs, we outline the different phases in the pilot, evaluation and subsequent migration to a customized NMT solution for a translation program of 28 languages.

3.1 Pilot Scope

For the pilot, we selected a subset of four languages out of the total 28. The selection of the languages was driven by several factors: 1) client priorities (translation volumes and cost) needed to be reflected, 2) we wanted to look at languages from different language families, 3) we had to stay within a fixed budget. Based on these parameters, German, French, Russian and Japanese were selected. We then went ahead with engine training in a commercially available, customizable NMT system. To ensure that results were comparable, the new NMT systems were trained with data identical to the data used for the existing SMT systems.

3.2 Evaluation Methodology

The setup of machine translation pilots is largely driven by client needs, the available budget, as well as the planned final program purpose and setup. Depending on this purpose and setup, one or more of the following options are usually selected to analyze the suitability and quality of a given machine translation engine:

- Automatic scoring: comparatively easy, quick and cost-effective analysis, thanks to our proprietary scoring tool; also the most common method for a quick comparison of different system builds and measuring quality on larger samples
- Human evaluation: a) for Utility to determine understandability for informational purposes only, b) for Adequacy/Fluency to get data on suitability for post-editing, c) in the form of an engine ranking of several engines, d) with error annotation to get a better picture on nature of errors per engine.
- Productivity testing: to get a picture of real post-editing performance, by measuring the time spent editing individual

[1] http://www.levenshtein.net/

sentences or averaged over larger documents, typically expressed as throughput in words per hour.

The long-term objective for the program in question was clearly defined: migrate an existing SMT post-editing program to NMT, in order to provide higher quality raw MT to post-editors, and eventually increase productivity and reduce cost. It was therefore important to include real productivity data in the pilot, more so than human evaluations and error annotations (at this stage).

For this particular pilot, we used the TAUS DQF Quality Dashboard[2], the related SDL Trados Studio plugin[3] and a proprietary analytics tool to capture throughput and productivity. Productivity was measured both on the customized SMT solution currently in place, and a customized NMT system, built with identical data.

Both translation and post-editing productivity, among other factors, largely depend on individual speed of the translator/post-editor. It is therefore recommended to use several resources for productivity tests and then average the results. For our pilot, we opted for two resources per language.

The decision to use the TAUS DQF Quality Dashboard and the related SDL Trados Studio plugin was driven by the following factors:

- Readiness due to existing company account with the Quality Dashboard
- Ease of use: SDL Trados Studio plugin enables fast and easy setup of test projects in the Quality Dashboard and Trados Studio.
- Known user interface: Testers can work in a familiar environment (Trados Studio), therefore their performance will not be affected by a new, unknown tool.

In addition to the productivity data, we also ran automatic scores on the completed translations for both custom SMT and custom NMT. As per our internal research over the past years, Edit Distance based on the Levenshtein algorithm seems to be one of the most useful automatic scores for comparing the quality of the raw MT for post-editing. It has turned out to be the most reliable metric in our evaluations as well as easily understandable for both translators and clients when shown in the form of a side-by-side comparison of edits (Marg et al., 2017a; Marg, 2016).

3.3 Pilot Take-Aways

Results from the pilot showed a clear productivity increase from customized NMT compared to the existing, customized SMT for German and Japanese, and lower, but still valid increases for French and Russian.

In contrast to the reliability of the Levenshtein Edit Distance in our evaluations over the past years, in the case of this pilot, Edit Distance results contradicted the increase in productivity for all languages but German. With Edit Distance being 3-6 percentage points higher from the customized NMT system for Japanese, French and Russian, this can be seen as a moderate difference, but still needs further research and investigation.

3.4 Next Steps

Based on the results of both the internal testing for various languages and content types (*generic NMT*, see 3 Planning for NMT for Enterprise-Scale Programs) and the client pilot for the selected languages (*customized NMT*), as well as general industry results, the client felt confident enough to go ahead and plan for a live rollout across 28 languages.

4 Migration

4.1 Assessment Criteria

When we selected the NMT provider for our client pilot, we made the decision based on the availability of customizable systems at that time, results from previous internal tests with this system, a good cooperation with the provider, the general customization options/ease of use, etc. After the completion of our pilot, other providers announced that they would release customizable NMT solutions later in 2018. To make sure to provide our client with the best option both technology- and cost-wise, we reevaluated the selection of the system to be used based on the following criteria:

- Customizable NMT readiness: later (other providers) vs. now (pilot provider)
- Connector to the existing Translation Management System (TMS): in place (other provider) vs. to be built (pilot provider)

[2] https://www.taus.net/quality-dashboard-lp
[3] https://www.taus.net/evaluate/dqf-plugin-for-sdl-trados-studio

- Customization options: What options for customization are exposed to the user? Is it possible, for example, to force client-specific terminology?
- Cost: Which of the available solutions would be more cost-effective overall?

For enterprise-scale translation programs, an automated workflow is essential. With several hundred to thousands of words processed per day and target language, manual file handling and injection of the machine translation output would simply not be manageable for project managers, both on client and on LSP side. This is where a TMS comes into play to:

- automate the injection of matches from the Translation Memory (TM), a database of previous translations, and
- automate the injection of machine translation, via an API connection to the MT system.

The development of such APIs or connectors between individual systems can be very costly and time-consuming. Therefore, using an MT system that already has a connector for the relevant TMS can decrease costs and time of deployment significantly. This would typically be the preferred option, provided this MT system is at least on par with systems that do not yet have such a connector (on par in relation to other decisive factors such as output quality and other costs). An existing API connection from our client's current TMS to their current SMT system was therefore the main reason to change the selection of the NMT system from the pilot provider to the client's existing SMT provider who would deploy customizable NMT later in 2018.

4.2 Rollout Plan

With the newly selected system, our NMT rollout plan had to factor in the following aspects:

- Languages available in generic NMT now + customizable as of release date
- Languages not available with NMT so far
- Current Edit Distance from existing SMT systems vs. Edit Distance from generic NMT now + anticipated Edit Distance with customizable version (all Levenshtein)

4.3 Challenges

Challenges during an early adoption enterprise-scale migration like the one described in this paper can be grouped into two categories:

- Availability of languages in the new system due to early adoption
- General migration challenges in relation to the involved technologies and processes

Due to the urgency of the planned migration, language availability and the resulting language migration sequence were the most pressing topics.

Out of the 28 languages to migrate for the program in question, 23 were available with generic NMT in the selected system—and were planned to be available as a customizable version later in 2018. 5 were not available with NMT at all and had to stay in the current customized SMT until this would change.

To potentially bridge the gap until customized NMT would become available, we decided to reevaluate the results from our internal tests with generic NMT. We scheduled an extended autoscoring comparison of the current customized SMT engines and generic NMT from the selected system for all 23 languages available with NMT thus far. We then came up with a definition of language groups based on their results from this comparison to determine which languages could potentially be moved to generic NMT prior to customization.

When it comes to general migration challenges, we first had to clarify whether the existing TMS would allow us to select different NMT systems (generic for some, custom for other languages). Additionally, as the MT provided by us is not only being used for post-editing by our own supply chain, but also that of other LSPs, changes in setup have to be communicated and managed with those LSPs to ensure continued stability for our end client. Finally, we would have to plan for additional post-editor trainings to help our supply chain with the change from SMT to NMT. Similar to publications by Burchardt et al. (2017) and Castilho et al. (2017), our evaluations had highlighted differences in the types of errors found in NMT and SMT output which would have an impact on the post-editing approach. While more analyses are required, it is important that the differences in error typology are communicated to all translation providers, to enable them to develop efficient methods and to

address all errors to the required final translation quality.

4.4 Research Proposal and Conclusion

During our session at the 21st Annual Conference of the European Machine Translation Association (EAMT 2018), we would like to present initial findings from this early adoption migration to NMT on an enterprise scale. We would like to demonstrate the solutions we implemented for the challenges outlined above, share details on the language migration sequence established based on our test results, and outline what additional challenges we might have come across during the migration.

5 Acknowledgement

We would like to thank our colleagues Elaine O'Curran, Alex Yanishevsky, Naoko Miyazaki and David Landan for their contribution.

References

Bahdanau, Dzmitry, Kyunghyun Cho and Yoshua Bengio. 2016-05-19. *Neural Machine Translation by Jointly Learning to Align and Translate.* Accepted as oral presentation at the 2015 International Conference on Learning Representations (ICLR 2015). arXiv:1409.0473v7 [cs.CL]. Accessed 26 March 2018

Burchardt, Aljoscha, Vivien Macketanz, Jon Dehdari, Georg Heigold, Jan-Thorsten Peter and Philip Williams. 2017. *A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines.* In The Prague Bulletin of Mathematical Linguistics (PBML), number 108, pages 159-170. https://ufal.mff.cuni.cz/pbml/108/art-burchardt-macketanz-dehdari-heigold-peter-williams.pdf Accessed 29 March 2018

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley and Andy Way. 2017. Is Neural Machine Translation the New State of the Art? In The Prague Bulletin of Mathematical Linguistics (PBML), number 108, pages 109-120. https://ufal.mff.cuni.cz/pbml/108/art-castilho-moorkens-gaspari-tinsley-calixto-way.pdf Accessed 29 March 2018

Cho, Kyunghyun, Bart van Merrienboer, Dzmitry Bahdanau and Yoshua Bengio. 2014-10-07. *On the Properties of Neural Machine Translation: Encoder–Decoder Approaches.* In Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8). arXiv:1409.1259v2 [cs.CL]. Accessed 26 March 2018

Kalchbrenner, Nal and Philip Blunsom. 2013. *Recurrent Continuous Translation Models.* In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1700–1709. Association for Computational Linguistics. http://www.aclweb.org/anthology/D13-1176 Accessed 26 March 2018

Marg, Lena. 2016. *The Trials and Tribulations of Predicting Machine Translation Post-Editing Productivity.* Presented at the 2016 Language Resources Evaluation Conference (LREC). http://www.lrec-conf.org/proceedings/lrec2016/pdf/810_Paper.pdf Accessed 26 March 2018

Marg, Lena, Naoko Miyazaki, Elaine O'Curran and Tanja Schmidt. 2017. *Comparative Evaluation of NMT with Established SMT Programs.* In Proceedings of MT Summit XVI, Vol. 2: Users and Translators Track, pages 166-178. http://aamt.info/app-def/S-102/mtsummit/2017/conference-proceedings/ Accessed 26 March 2018

Marg, Lena, Naoko Miyazaki, Elaine O'Curran and Tanja Schmidt. 2017. *Generic NMT vs. Established SMT—An Assessment in Relation to Post-Editing.* In 2017 School of Advanced Technologies for Translators (SATT) Teaching Material (available upon request from satt-2017@fbk.eu).

Sutskever, Ilya, Oriol Vinyals and Quoc V. Le. 2014-12-14. *Sequence to Sequence Learning with Neural Networks.* In Proceedings of Advances in Neural Information Processing Systems 27 (NIPS 2014). arXiv:1409.3215v3 [cs.CL]. Accessed 26 March 2018

https://www.taus.net/think-tank/news/press-release/dqf-and-mqm-harmonized-to-create-an-industry-wide-quality-standard Accessed 26 March 2018

https://www.taus.net/evaluate/dqf-plugin-for-sdl-trados-studio Accessed 26 March 2018

http://www.levenshtein.net/ Accessed 26 March 2018

https://www.taus.net/quality-dashboard-lp Accessed 26 March 2018