A Comparison of Statistical and Neural MT in a Multi-Product and Multilingual Software Company - User Study

Nander Speerstra
Machine Translation Researcher, Infor
Baron van Nagellstraat 89
Barneveld, Netherlands
Nander.Speerstra@infor.com

Abstract

Over the last 4 years, Infor has been implementing machine translation (MT) in its translation process. In this paper, the results of both statistical and neural MT projects are provide to give an insight in the advantages and disadvantages of MT use in a large company. We also offer a look into the future of MT within our company and to strengthen the implementation of MT in our translation process.

1 Introduction

In the last few years, we have seen a change of direction regarding machine translation approaches. In different domains, more research is being focussed on neural machine translation (NMT) in comparison to phrase-based statistical machine translation: in both the research environment (Bojar et al., 2016) and commercial companies like Google (Wu et al., 2016) and Microsoft (Awadalla et al., 2018) NMT is increasingly important.

In the context of commercial translations, the continuous improvement of (N)MT has not passed unnoticed. More and more language service providers (LSPs) are implementing machine translation into their translation workflows and in addition, translation teams in large companies are investing in machine translation as part of their translation processes.

As a large global software development company, Infor[1] translates its products into many languages. This paper summarizes the results of the investigations into the potential benefits of machine translation for a company with many products, many target languages and very different translation circumstances per product. This study consists of two main parts: SMT and NMT. First, we give a description of our experiments, after which the results of the experiments are described. Lastly, the results and impact on our company are discussed.

We had 2 main goals for this user study: to find out the current importance of (S)MT in our company and the potential benefits of moving to NMT in the future. These goals are discussed in Section 4.

2 Experiments

2.1 Background

Infor is an enterprise software company that currently markets more than 125 different products, translating any number of these into 49 separate languages. The translation process involves both internal translators (up to 15 languages) and LSPs. A visual representation of the MT workflow is presented in Figure 1. Once the documentation is finished by technical writers, the translatable files are pre-processed: sentences that have been translated in previous versions of the product are re-used to prevent re-translation of already translated content. Subsequently, machine translation and an automatic post editing script is run to fix some of SMT's errors. From here, the post-editing and translation are done by vendors or internal linguists, who also perform a quality check.

For many products, both the user interface and the documentation are translated into different languages. The documentation is written as online help and generally consists of relatively short sen-

[1] http://www.infor.com/

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. 315–321
Alacant, Spain, May 2018.

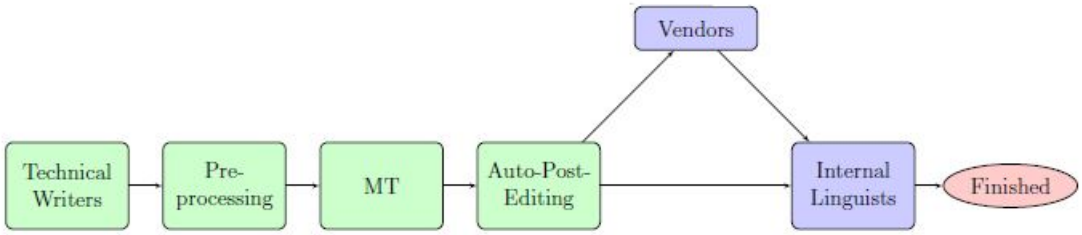

Figure 1: Translation workflow within Infor for MT projects

tences (1-15 words) with formatting and other tags. An example of this documentation material (English to Dutch) is given in Figure 2. User interface sentences often contain only one or very few words, which makes translation more difficult: often, different translations fit due to the shortness of the sentences while only one translation is terminologically correct.

The frequency of a product translation cycle varies: depending on the product, translation of edited existing and additional new materials may occur once, twice or twelve times per year. In addition, the number of times a product has been translated before (i.e. the amount of available training data) differs significantly: some products do not have a previous translation, others have been translated for over 20 years to certain languages.

As an example, the size of machine translation projects for three official Infor products (Infor LN, Infor BI and Infor d/EPM) is given in Table 1. The number of machine translated words differs per translation project, as does the update frequency.

Table 1: Number of machine translated words of 3 recent MT projects

Product	# words	# languages	Update cycle
LN	77,726	5	semi-annually
BI	109,922	7	annually
d/EPM	306,331	8	semi-annually

Most of Infor's documentation is written in US English and MT tests have only been performed on projects with English as the source language.

2.2 Statistical machine translation

Since 2014, MT projects have been executed at Infor using a moses-based statistical machine translation system from Morphologic Localisation: Globalese[2]. A handful of documentation translation projects were chosen as test projects for integrating MT in the translation process. These MT projects shared the following characteristics:

- They contained enough machine translatable segments to be worthwile
- There was sufficient training data (at least 50,000 sentences)
- Only some target languages were chosen of which most were close to the source language (English)

Two products were recurring to be machine translated for each occurring product update: LN and BI. The results of these product translations over the past two years (2016 and 2017) are discussed in Section 3.1.

For each of the products, one SMT system was used per language pair; i.e., if a product was translated to 6 languages, 6 SMT systems were trained and used for translation. This reflects the use of MT within Infor: we currently use one SMT system per product per language pair, as we do not generate parallel translations.

During the first tests, we noticed that MT makes a specific set of mistakes - often different mistakes per language. Therefore an automatic post-editing (APE) script was created that fixed basic errors introduced by the system, especially concerning tags. Example: 'Click on the <name> button' was machine translated to Dutch as

Druk op de <name> knop

while the following translation would have been correct:

Druk op de knop <name>.

APE fixes were only created for languages close to the source language (English), because the fixes required language-specific knowledge.

[2]http://www.globalese-mt.com/

Figure 2: Example of Infor documentation, product LN (English to Dutch)

While we added MT to the translation workflow for the above mentioned MT projects, we also ran tests on other products in order to find out if we could use MT for projects with:

- User Interface translations
- A low number of training segments
- Languages that are not closely related to English

These last tests were evaluated based on the expert opinions of our internal linguists and are not based on statistics. The reason for this is that the currently used evaluation metrics like BLEU and NIST correlate poorly with human judgment (Wang and Merlo, 2016), and our linguists have to work with the MT output: their opinions outweigh the statistical outcomes when a decision is made about using MT in translation projects.

Each of the SMT projects was set up with a quality threshold[3] and only segments with a quality estimation score of over 85% were retained, because sentences with lower scores were found to be sufficiently lacking in quality as to render them unusable. We selected this threshold after an evaluation of a first set of projects.

The results of these tests are shown in Section 3.1.

2.3 Neural machine translation

In the last few years, NMT has been the main interest in the machine translation industry. Globalese has recently released Globalese 3[4], a neural machine translation system which has subsequently been tested extensively at Infor. NMT is supposed to have several advantages over SMT. First, we explored the advantages of NMT. Then, we focused on tests using Globalese 3.

[3]https://web.archive.org/web/20150209082134/http://www.globalese-mt.com/product/features/quality-estimation/

[4]http://www.globalese-mt.com/2017/09/05/globalese-3-0-released/

The differences between SMT and NMT systems have been researched in depth and Jean et al. (2014) discuss several advantages of NMT. First, NMT requires very little domain knowledge. Where SMT requires a language model, NMT does not assume any linguistic characteristics and simply reads the source and target sentences as is. Moreover, an NMT model is trained as a whole, whereas an SMT engine consists of several separately trained parts including but not limited to (one or more) phrase table(s) and a language model. NMT also uses less memory than SMT systems that need to process large tables containing sentence pairs. Lastly, research has shown that NMT is more fluent and more accurate regarding word order (Toral and Sánchez-Cartagena, 2017).

Some of the disadvantages are discussed by Wu et al. (2016). The models need more training time than SMT models, NMT has difficulties with rare words and sometimes it translates sentences syntactically incorrectly. Also, long sentences are more often translated poorly by NMT (Toral and Sánchez-Cartagena, 2017).

For our company, some of the disadvantages appear to be less relevant since Infor's documentation contains very domain-specific terminology and rare words are not used frequently. Also, sentences are often relatively short. However, problems like an increased training time do matter: with many products and many languages to translate to, more training time could require a larger investment in resources.

One of our main questions is regarding the number of viable target languages. For SMT, we found that only languages related to English (Romance and Germanic languages) result in workable machine translations. Will NMT enable us to translate into additional languages, as Microsoft claims its new NMT system does with Chinese (Awadalla et al., 2018)?

As of Globalese 3.1, it is possible to use *core*

and *auxiliary* corpora as training data[5]. This core function makes sure that the core vocabulary is not overruled by the larger auxiliary corpora and, at the end of the training phase, the engine is further tuned to the core corpus. We created a test for Dutch, German and Russian, where an older BI project was selected to be re-translated with newly set up NMT engines. For each language, the translatable segments were processed with the following three machine translation systems:

- SMT
- NMT
- NMT with core functionality

The engines (SMT, NMT and NMT with core functionality) were trained using the number of training segments shown in Table 2. For this test the aforementioned SMT quality threshold of 85% was removed because the NMT systems from Globalese did not have a quality estimation script with which to compare. The test files for all engines were pre-translated as usual and the remaining 7203 sentences (77,261 words) were machine translated. These sentences were evaluated by internal linguists (one linguist per language).

Table 2: SMT vs. NMT: Translation project training size for Dutch, German and Russian

Language	# training segments
Dutch	499,106
German	275,887
Russian	198,360

This test includes two of our main questions: do we need more data with NMT than with SMT (i.e. will Russian and German be evaluated with worse results for NMT than for SMT) and can we translate to more languages without quality loss (i.e. are the evaluations for Russian similar to those for Dutch and German)? The three sets of translated files were given to internal linguists for evaluation without information on the engines that were used to produce them.

3 Results

Normally, machine translation results are expressed using evaluation scores like METEOR, BLEU and/or hTER. However, as these metrics generally do not correlate with linguists' findings (Sun, 2010), we chose to only report the number of machine translated segments (that were used in the translation projects) and the qualitative analyses of our linguists. Both the linguist reviews and the number of machine translated sentences gave us an indication of the usefulness of MT in translation projects.

3.1 Statistical machine translation

In the period 2016-2017, roughly 900,000 words have been machine translated using SMT for the products Infor LN and Infor BI. In Table 3, the number of translated words is shown for the last 2 years. The decreased number of machine translated words for BI in 2017 is caused by changes in the MT setup as a result of an evaluation of the 2016 results. These changes are discussed in Section 4.1.

Table 3: Number of SMT words for 2 products, in the period 2016-2017

Product	2016	2017	Total
LN	285,857	292,095	577,952
BI	259,530	56,174	315,704
Total	545,387	348,269	893,656

SMT was found to be useful in the translation projects of 10 products with a total of 2,026,760 machine translated words. In the largest MT project (BI 2016), translations were run from English to 12 different languages: Brazilian Portuguese, Danish, Dutch, French, German, Italian, Japanese, Norwegian (Bokmål), Russian, Simplified Chinese, Spanish and Swedish.

For three tests, the quality of the translations was insufficient for use in actual translation: tests of user interface translations, projects with a low amount of training segments and target languages that are not closely related to English. The user interface translations contained sentences that were too short and ambiguous for MT, which often led to incorrect translations. Projects with a low number of training data often resulted in very few workable translations due to the quality estimation threshold of 85%. Unrelated target languages resulted in poor translations and were not selected for new translation projects.

We did not have statistical metrics for the MT projects, but the discount on MT words is an indication of the importance of MT. For the project

[5]http://www.globalese-mt.com/2017/10/31/augmented-in-domain-engines/

BI 2016, we were given an average discount of 67% on machine translated sentences on an average word price of 15 ct/w. To that extent, the BI 2016 project led to a cost saving of €25,953.

3.2 Neural machine translation

Besides the motivations for using NMT over SMT in the literature,we performed a qualitative analysis on 3 sets of translations of the product Infor BI: translations using SMT, NMT and NMT with the core functionality. Internal linguists, one per language, were asked to rank the quality of the translation sets and give examples of correct and incorrect translations. Each of them returned the following ranking: (1) NMT with core functionality, (2) NMT and (3) SMT. The quality of (1) and (2) was comparable but with a slight preference for (1), (3) was said to have less workable translations compared to (1) and (2). This was expected for Dutch and German as we had enough training data for those languages, but also our Russian team evaluated NMT as more useful than SMT. The linguist for Dutch mentioned the quality of NMT with core functionality as follows: 'I think this version of the project is very good and MT is a great time saver here, not only because post editing doesn't seem so strenuous.'

For all languages, the results can be summarized as follows. SMT had many different issues, from incorrect word/tag order, incorrect capitalization, incorrect word order to illogical translations. Although most issues are minor, they were too numerous to make the translations directly usable and required heavy post-editing.

NMT and NMT with core functionality also had difficulties with word/tag order and word order in general. And, in contrast with SMT, NMT made strange (albeit fluent) semantic errors, where the translation was incomprehensible. An example of such an NMT error is shown in Figure 3, together with examples of errors concerning text in tags and word omissions. But compared to SMT, NMT was said to contain more workable translations and would take less post-editing time. Short sentences especially were much more often correct.

Consequences of this test will be discussed in Section 4.2.

4 Discussion

In this section, the results of the SMT and NMT experiments are discussed.

4.1 Statistical machine translation

As described in Section 3.1, about 2 million sentences have been machine translated with our SMT engines in the period 2014-2017. There are several points of interest that need a more elaborate discussion: the output quality, the number of languages found workable for SMT and the project initiation time.

4.1.1 Output quality

Overall, the output quality was good enough to use MT in translation projects. As this was a goal of machine translation (decreasing costs by post-editing instead of translating from scratch), SMT has been successfully used in translation projects. Because of the 85% threshold in official projects, about 40-50% of the translatable segments were actually machine translated. Increasing the quality of the output (and thus increasing the number of machine translated segments) is one of the key research areas within our company, as this affects the costs of translation projects directly.

4.1.2 Number of languages

During our experiments, we found that target languages close to the source language were translated with a higher quality than target languages outside of the Romance and Germanic families. Since our projects have English as the source language, Germanic and Romance languages were most suitable for machine translation. Early tests on Chinese (zh-CN) and Japanese showed that, to our standards, those languages resulted in a quality unsuitable for use in actual projects.

Another issue with SMT was the necessity of an automatic post-editing script. This script fixed some known issues for specific languages, but this could only be set up by language experts. As our team does not have expertise in languages outside the Germanic and Romance families, only these languages had APE scripts.

4.1.3 Project initiation time

Because SMT requires several individual components to be trained, re-training the engines for a translation project was sometimes rather time-consuming. Especially when the number of languages in a project was high, it took several hours to manually prepare the engines. Although some actions were scripted, uploading new training segments and creating engines was at the time of the

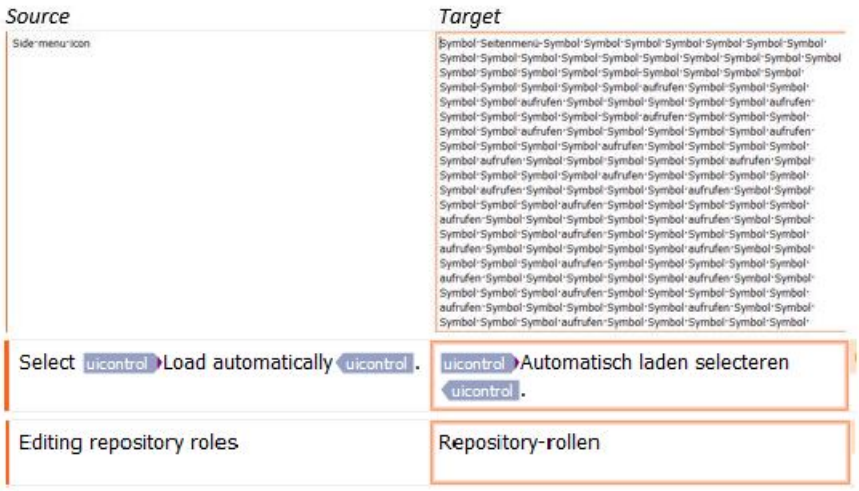

Figure 3: Examples of NMT errors in our test project: hallucination translations, words incorrectly placed in tags and omission of words

test project not yet available. We updated the engines after every project to make sure that the engines are trained on as much data as possible.

4.2 Neural machine translation

The tests using our neural engines have given a useful insight in the advantages and disadvantages of neural machine translation. In this section, the advantages and disadvantages are discussed. In both sections, a link is made to our SMT results.

4.2.1 Advantages

First of all, our test on Dutch, German and Russian showed that for all 3 languages (1) the NMT quality resulted in workable translations and (2) NMT is preferred over SMT. Where (1) was expected for the closely related language pairs English-Dutch and English-German, we weren't certain for Russian: in our SMT projects, the Russian language appeared to be too different from English to obtain workable translations. But the experiment showed that NMT resulted in useful output for Russian as well. Secondly (2), NMT was preferred over SMT for every language pair.

Another advantage of NMT is the time gain when preparing the engines. Because only 1 model was needed per language (compared to the multiple components in an SMT engine), the preparation time was significantly lower: for the NMT models in the test, preparation took only 10 minutes instead of the 30 minutes that it took to set up the SMT engines.

4.2.2 Disadvantages and solutions

Our NMT tests also revealed some of the downsides of NMT: training the engines took much longer than with SMT (2-3 times longer), more data was needed and the output was sometimes less reliable.

The training time appeared to be problematic at first, because we re-trained the engines before each project. This would require more resources with NMT with the same (or more) languages per product. Given the product LN (6 engines, one for each package, and 6 languages) and a training time of 2 days, this would result in a total semi-annual training time of 72 days (3 months on every 6 months).

However, we have not investigated whether it is necessary to update the engines after each translation project. If we would only update once a year, the effect on our resources is reduced.

Furthermore, we needed more data. Although not presented here, we have translation projects with less than 50,000 segments as training data. SMT was capable of generating qualitative output (for closely related language pairs), NMT was not. However, due to the core functionality function, we have been able to merge data from several projects into one large engine without causing terminology issues. This is a major improvement, as we can potentially machine translate each product for which we have enough training data in that specific language. This would also decrease the necessity of re-training engines after each project, because the new set of translations would have less impact in the large engines.

Lastly, the output was less reliable. Although

NMT is more fluent (Skadina and Pinnis, 2017), the output is less accurate and can sometimes miss the point completely. But this has been found to be an advantage by some of our linguists: because translations are more fluent than with SMT, it is easier to see that the translations should be removed and re-translated from scratch. This saves time when post-editing MT sentences.

5 Conclusion and outlook

In this paper, we have discussed the outcomes of statistical (SMT) and neural (NMT) machine translation experiments that we have conducted at Infor. With a total of over 2 million machine translated words, SMT has become a significant factor in product translations. SMT has been used for 10 products with up to 12 languages. Tests showed that SMT produced workable translations on language pairs that are closely related, and we needed handwritten auto-post-editing scripts to improve the output quality. A first test with NMT has shown that NMT performs better on all languages tested (Dutch, German and Russian) than SMT.

The purpose of the experiments was to determine the significance of MT in our workflow and whether NMT is the next step to take. Based on the number of machine translated words in the last few years, we now have a good understanding of the type of projects in which MT is of use, and it has already impacted the costs of translation projects in which MT was used. We have also seen that NMT scores higher than SMT according to our linguists, which is a clear indication that NMT is the next step in improving our MT process. With a potential of many more products to translate and many more languages to translate to, we will start experimenting with NMT in the same way that we did with SMT.

Acknowledgements: I would like to thank Gábor Bessenyei and Gergely Horváth from Globalese for giving us the opportunity to test their new neural machine translation system. I would also like to thank Thijs Trompenaars and my colleague John Musters for their important and useful reviews on this paper.

References

Awadalla, Hassan, Hany and Aue, Anthony and Chen, Chang and Chowdhary, Vishal and Clark, Jonathan and Federmann, Christian and Huang, Xuedong and Junczys-Dowmunt, Marcin and Lewis, Will and Li, Mu and Liu, Shujie and Liu, Tie-Yan and Luo, Renqian and Menezes, Arul and Qin, Tao and Seide, Frank and Tan, Xu and Tian, Fei and Wu, Lijun and Wu, Shuangzhi and Xia, Yingce and Zhang, Dongdong and Zhang, Zhirui and Zhou, Ming. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *In proceedings*.

Bojar, Ondrej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Jimeno Yepes, Antonio, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation (WMT16). *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers*.

Jean, Sébastien and Kyunghyun Cho and Roland Memisevic and Yoshua Bengio. 2014. On Using Very Large Target Vocabulary for Neural Machine Translation. CoRR, Vol. abs/1412.2007.

Skadina, Inguna and Marcis Pinnis. 2017. NMT or SMT: Case Study of a Narrow-domain English-Latvian Post-editing Project. *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers.*

Sun, Yanli. 2010. Mining the Correlation between Human and Automatic Evaluation at Sentence Level. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Toral, Antonio and Víctor M. Sánchez-Cartagena. 2017. A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. CoRR, Vol. abs/1701.02901.

Wang, Haozhou and Paola Merlo. 2016. Modifications of Machine Translation Evaluation Metrics by Using Word Embeddings. *Proceedings of the Sixth Workshop on Hybrid Approaches to Translation*, Osaka, Japan, December 2016, 33–41. *CoRR*.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, ukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR*.