Can Out-of-the-box NMT Beat a Domain-trained Moses on Technical Data?

Anne Beyer*, Vivien Macketanz△, Aljoscha Burchardt△ and Philip Williams◇
* beo Gesellschaft für Sprachen und Technologie mbH
Ruppmannstrae 33b, 70565 Stuttgart, Germany
anne.beyer@beo-doc.de
△ German Research Center for Artificial Intelligence (DFKI)
Language Technology Lab, Alt-Moabit 91c, 10559 Berlin, Germany
firstName.lastName@dfki.de
◇ University of Edinburgh
School of Informatics, 10 Crichton Street, Edinburgh, EH8 9AB, UK
pwillia4@inf.ed.ac.uk

Abstract

In the last year, we have seen a lot of evidence about the superiority of neural machine translation approaches (NMT) over phrase-based statistical approaches (PBMT). This trend has shown for the general domain at public competitions such as the WMT challenges as well as in the obvious quality increase in online translation services that have changed their technology. In this paper, we take the perspective of an LSP. The questions we want to answer with this study is if now is already the time to invest in the new technology. To answer this question, we have collected evidence as to whether an existing state-of-the-art NMT system for the general domain can already compete with a domain-trained and optimised Moses (PBMT) system or if it is maybe already better. As it is well known that automatic quality measures are not reliable for comparing the performance of different system types, we have performed a detailed manual evaluation based on a test suite of domain segments.

1 Introduction

In the last year, we have seen a lot of evidence about the superiority of neural machine translation approaches (NMT) over phrase-based statistical approaches (PBMT). This trend has shown for the general domain at public competitions such as the WMT challenges (Bojar et al., 2016) as well as in the obvious quality increase in online translation services that have changed their technology.[1]

When it comes to particular domains in the context of commercial translation services, the interest in NMT is huge, but we are not aware of systematic public studies about the performance of NMT in comparison to PBMT. While bigger companies are already in the process of changing their technology, smaller language service providers (LSP) have limited resources in their day-to-day-business both in terms of humans and compute power for undertaking the necessary experiments. For researchers, it is still difficult to obtain suitable training data in order to assess the potential of the new technology on in-domain data.

The background for this study was simply the question of an LSP if now is already the time to invest in the new technology. To answer this question, we wanted to collect evidence as to whether an existing state-of-the-art NMT system for the general domain can already compete with a domain-trained and optimised Moses (PBMT) system or if the former can maybe even outperform the latter already.

As we did not want to rely solely on automatic measures, we have performed a manual evaluation based on a phenomenon-driven test-suite, a method we have applied for evaluations in the technical domain before, e.g., in (Avramidis et al., 2016).

2 Experiment

2.1 Data

The customer data used in this study came from translations of catalogues for technical tools. Our

[1] https://research.googleblog.com/2016/09/a-neural-network-for-machine.html

dataset consisted of translation tasks from German into British English assigned to beo over a course of two months. Overall, the set contained around 5,000 segments.

2.2 Phrase-based Statistical MT System

The PBMT system used is based on Moses (Koehn et al., 2007) and was adapted to integrate MT into the translation workflow at beo.

As training data we used the customer's translation memory (TM) and terminology, which yielded a total of 337,600 segments. Formatting tags were removed from the data and it was tokenized and lower cased. As we translate from German, compounds were also split on the source side in order to reduce data sparsness in terms of unknown words. A 3-gram language model was built using IRSTLM (Federico et al., 2008).

The training procedure follows the baseline Moses setup[2], but the model was not tuned further, as no tuning setup was found yet which improved the system's performance over the baseline, according to an internal evaluation with our translators. This is similar to what we found for other customer set-ups. It could be due to the fact that the training-data and the translations are very similar, as we only used in-domain data for training. We have not yet tried to add more out-of-domain data because this did not improve the usefulness of systems trained for other customers, but might look into that at a later point as well. As we are only concerned with the application of MT for post-editing, the quality requirements are different from other tasks such as quality evaluation and we rely more on post-editor feedback that automated quality scores.

For the translation, we used the M4Loc integration tools[3], a wrapper for Moses which extracts formatting tags before the translation and inserts them into the target afterwards according to the word alignnment (Hudk and Ruopp, 2011). Furthermore, we ran a few test rounds on the customer data together with our translators and created a set of hand-crafted rules based on regular expressions which are applied after the MT to fix certain errors (e.g., with casing or spaces).

[2] http://www.statmt.org/moses/?n=Moses.Baseline

[3] https://github.com/achimr/m4loc

2.3 Neural MT System

The neural system that was used in this study was built by the University of Edinburgh. This MT engine is the top-ranked system that was submitted to the WMT '16 news translation task (Sennrich et al., 2016). The system was built using the Nematus toolkit.[4]

As training data, only the official WMT task data was used – this system did not have access to the customer-specific data during training. The data was tokenized and truecased, and tokens on both the English and German sides were split into subword units using byte-pair encoding (BPE), a frequency-based method that aims to improve the handling of rare words.

The full training configuration and scripts for this system have been publicly released.[5]

2.4 Manual Evaluation Procedure

For the manual evaluation process, two professional (computational) linguists went through the data and identified reoccurring linguistic phenomena that are characteristic for this domain-specific data.[6] In a second step, all the phenomena detected were narrowed down to the most prominent ones, namely formal address, genitive, modal construction, negation, passive voice, predicate adjective, prepositional phrase, terminology and tagging. Thereafter, 100 segments per phenomenon were extracted, resulting in a total of 900 segments. For each segment, the total occurrences of the respective phenomenon were counted. Then, the total occurrences of the phenomena in the MT outputs were counted. Consequential, translation accuracy was calculated by dividing the number of occurrences in the MT output by the total number of occurrences in the segments.

When evaluating the correctness of the translations, the focus lies solely on the respective phenomenon under consideration, other errors are ignored. For a translated phenomenon to be counted as correct, it does not necessarily exactly have to match the reference, but it can also be realized in a different linguistic construction expressing the same semantic meaning, e.g., a passive construction that is translated in active construction will

[4] https://github.com/rsennrich/nematus

[5] https://github.com/rsennrich/wmt16-scripts

[6] These "linguistic phenomena" are understood in a pragmatic sense and include a wide range of issues that can influence the translation quality.

have less components but if the meaning is translated correctly, the counting should be adjusted to the instances in the source accordingly.

3 Evaluation Results

Due to the repetitive nature of the customer data, some of the segments in our dataset were already part of the TM or very similar to segments in the TM and therefore part of the training data for the Moses system. In order not to distort the results too much, those segments where Moses exactly matched the reference translations were omitted from the automatic evaluation. For the manual evaluation, we did not exclude those segments.

3.1 Automatic Evaluation Results

Even though BLEU is not intended to be used in order to compare different MT systems, this is a practice that is performed quite often. In order to show how much different translation quality evaluation methods can vary, we also carried out an evaluation on BLEU and METEOR, cf. Table 1. For calculating the automatic score, all tags were removed from the segments and the reference, furthermore all numbers were replaced by "10" because there were cases in which the reference involved different tags/numbers than the segments.

	NMT	Moses
BLEU	23.68	47.98
METEOR	28.46	38.26

Table 1: BLEU and METEOR scores.

As described above, the automatic evaluation has a clear bias towards Moses. This is amplified by the fact that the references were derived from post-edits of the Moses output. These segments are thus naturally more similar to the Moses output than to the completely independent NMT output. Despite removing the segments for which the translation by Moses exactly matched the reference, both BLEU and METEOR show distinctly better scores for Moses compared to the NMT system. Taking into account the manual evaluation, though, gives a different picture.

3.2 Manual Evaluation Results and Examples

Table 2 shows the results of the manual evaluation on segment-level. For the 900 segments extracted, 1,453 phenomena could be found altogether, as there was often more than one occurrence of the phenomenon per segment. Phenomena like terminology occur more frequently than phenomena like negation that rarely appear more than once within one segment. Percentage values in boldface indicate that the systems is significantly better on the respective phenomenon with a 0.95 confidence level.

	#	NMT	Moses
formal address	138	90%	86%
genitive	114	**92%**	68%
modal construction	290	**94%**	75%
negation	101	93%	86%
passive voice	109	**83%**	40%
predicate adjective	122	81%	75%
prepositional phrase	104	81%	75%
terminology	330	35%	**68%**
tagging	145	83%	**100%**
sum	1453		
average		**89%**	73%

Table 2: Manual evaluation translation accuracy focusing on particular phenomena.

The NMT system outperforms Moses on three categories: genitive, modal construction and passive voice. Moses on the other hand outperforms NMT on terminology and tagging – which is not surprising as terminology was part of the TM and tagging was handled by an extra module. For the remaining phenomena, the systems show no statistical significantly variance. Additionally, the NMT system also outperforms Moses on the overall average.[7] Nevertheless, it is important to keep in mind that the values of the manual evaluation only give insights on certain phenomena and do not necessarily represent the systems' overall performance but can rather be interpreted as revealing a tendency. Interestingly, the tendency the manual evaluation displays is counter to that of the automatic scores shown in Table 1. This can be traced back to the training material for Moses which included the the customer's translation memory and terminology which has a high influence on the BLEU and METEOR scores. The manual evaluation results on the other hand imply that even if a translation deviates substantially from a given reference it can

[7] Average calculation: division of the sum of the absolute numbers of correct segments by the sum of all segments for each system.

still be correct, a fact that is not taken into account in the automatic scores.

The following examples depict interesting findings from the analysis and comparison of the two systems. The relevant component of the sentence is underlined. When a system created a correct output for the respective phenomenon, the system name is marked in boldface.

(1) Source: Schweißbänder erhöhen wesentlich den Tragekomfort eines Helmes.
Ref.: Sweatbands significantly increase the wearing comfort of a helmet.
NMT: Welding tapes significantly increase the comfort of a helmet.
Moses: Welding belts significantly increase the wearing comfort of a Helmes.

Example (1) contains the genitive *eines Helmes* that should correctly be translated as *of a helmet*. As can be seen, the NMT correctly translates the genitive while Moses leaves *Helmes* untranslated which makes it hard to tell whether it correctly translates the genitive. This was a systematic problem for Moses, as Moses left unknown words untranslated. The NMT system on the other hand often generated sentences that were grammatical and contained "only" mistranslated unknown words rather than untranslated unknown words. As a result, syntactic features like the genitive in example (1) can be maintained.

(2) Source: Dazu kann das Board werkzeuglos gedreht und wieder eingehängt werden.
Ref.: The board can be turned and reattached without using tools.
NMT: The board can be rotated and remounted.
Moses: To do this, the board can be rotated and back.

Example (2) includes a modal verb construction. A modal verb is always followed by at least one other verb. In the construction above, the modal verb *kann* is followed by the two verbs *gedreht* and *eingehängt* as well as the verb *werden*. Those verbs form a processual passive construction. In order to count as correctly translated, the English MT outputs should also exhibit four verbs, as the construction is formed the same way in English. While the NMT system correctly translated all four verbs, Moses leaves out one verb. Note that the fact that both systems do not translate *werkzeuglos* (*without using tools* in the reference) can be ignored in this evaluation as the focus lies exclusively on the phenomenon of modal verb constructions.

(3) Source: Die Panoramascheibe mit integriertem Seitenschutz sorgt für eine <g id="1004">optimale Augenraumabdeckung</g>.
Ref.: The panoramic lens with integral side protection ensures <g id="1004"> optimum coverage of the eye area </g>.
NMT: The panorama disc with integrated side protection ensures a <g id="1004"> optimal eye room cover </g>.
Moses: The panoramic lens with integral side protection ensures <g id="1004"> optimum Augenraumabdeckung </g>.

The third example given here is taken from the terminology category. Additionally, it contains tagging which can be ignored in this case. The source sentence contains three terms: *Panoramascheibe*, *Seitenschutz* and *Augenraumabdeckung* which should be translated as *panoramic lens*, *side protection* and *coverage of the eye area*, respectively. The NMT system only correctly translates *side protection* while mistranslating the other two terms, giving literal translations. Moses correctly translates two of the three terms, leaving *Augenraumabdeckung* untranslated. Nevertheless, at first glance the NMT output looks "better" because it does not leave words untranslated. When taking a closer look though, this assumption does not hold.

As Moses benefits in terms of knowing a subset of the terminology, we considered it reasonable to also analyze segments without terminology in order to draw some more general conclusions about the comparison between the two systems, independent of the domain. For this purpose, 90 segments without domain-specific terminology were extracted from the data set. These segments comprise 30 short ($<$ 40 characters), 30 medium-length (40 - 79 characters) and 30 long ($>$ 79 characters) items. Two annotators were asked to evaluate these segments individually, rating them on a scale from 1 - 3, with 1 = perfect translation, 2 = small errors, content still understandable, and 3 = unintelligible. The mean values of the two annotators can be found in Table 3. While the NMT's

performance is judged better for the longer segments, Moses' performance is judged better for short and medium-length segments. Nevertheless, conducting a *t*-test showed that the differences in the mean values are not statistically significant. Yet, it should be kept in mind at this point that we did not expect the differences to be statistically significant as the population of segments examined was very small. We interpret the scores solely as a tendency.

Below, we will discuss an example from this category:

(4) Source: Neben den Bedingungen zur Aufstellung und Inbetriebnahme wird eine Vielzahl von technischen und gesetzlichen Anforderungen an das Lager selbst gestellt, um z. B. wassergefährdende Flüssigkeiten, Säuren und Laugen oder auch entzündbare Flüssigkeiten gesetzeskonform aufzubewahren und zu lagern.

Ref.: In addition to the conditions for erection and commissioning there are a wide variety of technical and legal requirements on the storage location itself, relating for instance to water-polluting liquids, acids and alkalis or also flammable liquids, which must be kept safe and stored in accordance with regulations.

NMT: In addition to the conditions for installation and commissioning, a wide range of technical and legal requirements will be placed on the warehouse itself in order to maintain and store, for example, water-hazardous liquids, acids and foliage, or even flammable liquids.

Moses: In addition to the conditions for erection and commissioning is a wide variety of technical and legal requirements of the stored even, e. g. for water-polluting liquids, acids and alkalis or flammable liquids legally compliant aufzubewahren and storage.

Example (4) belongs to the long segments, having 293 characters. While there were long segments that consisted of several sentences, this segment comprises only one sentence. It contains an infinitive clause that reaches from the conjunction *um* to the verb *zu lagern*. While in German, objects are located between the conjunction and the last verb, in English the conjunction *in order to* is immediately followed by the verb in the infinitive with the objects being located behind the verb. The NMT system successfully manages to resolve this construction, placing the verbs at the right position while Moses not only leaves the verbs at the end of the sentence but also leaves one verb untranslated. This example depicts our finding that NMT can handle long sentences better than Moses.

	∅ NMT	∅ Moses
short segments	1.7	1.5
medium-length segments	2.1	1.9
long segments	2.2	2.3

Table 3: Mean values for segments without terminology.

At the same time, this sentence also highlights difficulties that can arise, e.g., for post-editing, by the fact that the NMT system substitutes unknown words in the source with similar words in order to be able to translate them. While in some cases this might work out well, there are other cases where it does not, as in example (4) above: The word *Laugen* (*alkalis*) was treated as the word *Laub* which means *foliage*, resulting in a rather curious translation. For post-editing this means that in order to detect erroneous translations it is crucial to check the NMT output very thoroughly because mistranslations might be harder to find than in a system output that contains untranslated words.

4 Conclusion and Outlook

From the viewpoint of the linguistic phenomena we have studied in our experiment, the answer to the question in the title of this paper would probably be a sentence beginning with "Yes, but . . . ". The reason for the restriction is that the two categories NMT can not yet handle as good as Moses are of high importance in the language business: tags and terminology.

Still, sooner rather than later there will be tag-handling components for NMT systems and the issues with terminology will probably vanish once the NMT is trained on customer domain data. So, from the analytic perspective we took here, NMT could indeed become a valid alternative to PBMT

for commercial use in the future.

The purpose of this study was to determine if now is already the time for LSPs to start investing in NMT. Our comparison showed that even an out-of-the-box system can perform quite reasonably, although it was not trained on the specific data. Our next step will be to look into the OpenNMT system[8] and to compare models trained on the same dataset. Here, we will also take a closer look at other important factors, such as the time and effort needed for setting up such a system, the different training and decoding times and the impact of different kinds of errors on the post-editing effort.

For this purpse, We plan to also perform productivity tests with post editors to get a second, less phenomenon-driven comparison between the systems. In this course, we may also re-calculate automatic scores using post-edits as reference translations to rule out the Moses bias we have clearly observed in the figures we have presented here. For scenarios without post-editing, it would also be interesting to repeat task-based evaluations like the one we present in (Gaudio et al., 2016).

Another follow-up study that could be conducted might focus on a comparison of systems which are more similar with regard to their setup of the training data. In doing so, it would be interesting to investigate whether, for instance, an NMT system's BLEU and METEOR scores might get closer to those of an SMT system, and if the bias towards the NMT system in the manual evaluation scores persists or even increases.

Acknowledgement

This article has received support from the EC's Horizon 2020 research and innovation programme under grant agreements no. 645452 (QT21). We thank the anonymous reviewers for their valuable feedback.

References

Avramidis, Eleftherios, Vivien Macketanz, Aljoscha Burchardt, Jindrich Helcl, and Hans Uszkoreit. 2016. Deeper Machine Translation and Evaluation for German. In Hajic, Jan, Gertjan van Noord, and Antnio Branco, editors, *Proceedings of the 2nd Deep Machine Translation Workshop. Deep Machine Translation Workshop (DMTW), October 21, Lisbon, Portugal*, pages 29–38. Charles University Prague, Charles University, Prague, 10.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, 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. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany, August. Association for Computational Linguistics.

Federico, Marcello, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *9th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1618–1621, Brisbane, Australia, September.

Gaudio, Rosa, Aljoscha Burchardt, and Antonio Branco. 2016. Evaluating Machine Translation in a Usage Scenario. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

Hudk, Tom and Achim Ruopp. 2011. The integration of moses into localization industry. In *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT)*, pages 47–53, Leuven, Belgium, May.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for WMT 16. *CoRR*, abs/1606.02891.

[8] `http://opennmt.net/`