Statistical vs. Neural Machine Translation: A Comparison of MTH and DeepL at Swiss Post's Language Service

Volkart, Lise; Bouillon, Pierrette; Girletti, Sabrina

Statistical vs. Neural Machine Translation: A Comparison of MTH and DeepL at Swiss Post's Language Service

Lise Volkart, Pierrette Bouillon, Sabrina Girletti
FTI/TIM University of Geneva
Boulevard du Pont-d'Arve 40
1211 Geneva, Switzerland
lise.volkart@unige.ch
pierrette.bouillon@unige.ch
sabrina.girletti@unige.ch

Abstract

This paper presents a study conducted in collaboration with Swiss Post's Language Service that aims to compare the performance of a generic neural machine translation system (DeepL) and a customised statistical machine translation system (Microsoft Translator Hub, MTH) in terms of post-editing effort and quality of the final translation for the language direction German-to-French. The results for automatic and human evaluations show that DeepL is overall better than MTH, but its quality is underestimated by the BLEU score.

1. Introduction

Machine translation (MT) seems to raise the interest of a growing number of actors in the translation field. Willing to integrate MT in its workflow, the Swiss Post's Language Service asked us to accompany them in this process (see also Bouillon et al., 2018) . This study aims to compare a customised Statistical MT system (SMT) (i.e. Microsoft Translator Hub MTH) with a generic neural MT system (NMT) (i.e. DeepL) for the language direction German-to-French, in order to provide an answer to the following question: *Can a generic neural system compete with a customised statistical MT system?* We also questioned the reliability of the automatic metric we used and introduced the following subsidiary question in our work: *is BLEU (Papineni et al., 2002) a suitable metric for the evaluation of NMT?*

The paper is structured as follows: Section 2 briefly presents the customisation of MTH and the selection of the best performing system. In Section 3, we present how we compared MTH and DeepL by performing an automatic evaluation, a post-editing productivity test and a comparative evaluation of post-editing (PE) results. In Section 4, we address our subsidiary research question by presenting the correlation between the human and automatic evaluations. We conclude our work in Section 5.

2. Customisation of MTH

Four translation memories (TMs) (288,211 segments in total) and four domain–specific glossaries (2,217 terms in total) provided by the Swiss Post's Language Service were used to train several systems with the MTH platform for the four Swiss Post subject areas: *vocational training*, *financial services*, *process manual* and *annual report* (Volkart, 2018). These systems were evaluated using the automatic metric BLEU (Papineni et al., 2002). The best system was obtained for the subject area *annual report* (Bleu = 41.36) with all the four TMs and 76 domain-specific glossary terms.

Proceedings of the 40th Conference Translating and the Computer, pages 145–150,
London, UK, November 15-16, 2018.

3. Comparison of MTH and DeepL

To answer our first research question, we compared the output produced by our best MTH system on the subject area *annual report* with the output produced by DeepL on the same domain. To do so, we conducted both an automatic evaluation and a human evaluation. The automatic evaluation was done with BLEU (see Section 3.1). For the human evaluation, we decided to undertake a task-based human evaluation in the form of a post-editing (PE) productivity test as the Swiss Post is interested in using MT as a pre-translation tool (see Section 3.2). As the Swiss Post is interested in increasing the productivity of its translators while keeping a high level of quality in the translation, we also conducted a second human evaluation to assess the quality of the output of each system after PE (see Section 3.3).

3.1. Automatic evaluation

For this evaluation, a test set containing 1718 unseen segments was created by exporting the newly added segments from the *annual report* TM. We translated this test set both with our best MTH system and DeepL, then we calculated the BLEU score obtained by each system. The results are presented in Table 1 and show similar scores for both systems.

System	BLEU
DeepL	25.23
MTH	23.46

Table 1: BLEU scores obtained by MTH and DeepL.

3.2. Post-editing productivity test

For this evaluation, a subset of 250 segments was randomly extracted from the test set used for the previous automatic evaluation. We translated this test set using both the best MTH system and DeepL.

Two translators (one in-house translator from the Swiss Post's Language Service and a freelance translator) participated in this evaluation. We asked them to post-edit the output produced by the two systems (500 segments in total). The source-target segment pairs for each system had been mixed in such a way that the evaluators would not know the origin of the output (i.e. which system produced the translation) and would never post-edit two identical segments in a row. The post-editing task was performed on the platform MateCat[1], which records the PE time for each segment and for the whole project. Each evaluator was given a brief introduction on post-editing[2] before the experiment and was asked to perform a full PE (i.e. to post-edit the output in order to obtain a quality comparable to a human translation) following TAUS' guidelines (TAUS et CNGL, 2010).

3.2.1. Results

We compiled the PE time for both systems and both evaluators. In order to compare both systems in terms of the amount of corrections made by the post-editors, we compiled the HTER score (Snover et al., 2006). These two measures (time and HTER) combined gave us

[1] https://www.matecat.com

[2] The information and guidelines given to the evaluators are given in detail in Volkart (2018).

an idea of the PE effort on the part of the post-editors. The PE time and HTER obtained are presented in Table 2.

System	Post-editor	HTER	Time (s/word)
MTH	Post-editor 1	0.5044	4.6
	Post-editor 2	0.4639	9.94
	Average	0.4842	7.27
DeepL	Post-editor 1	0.1627	2.57
	Post-editor 2	0.0780	4.18
	Average	0.1204	3.38

Table 2: HTER scores and average post-editing time needed per word (in seconds per word).

Both evaluators were much faster when post-editing the output produced by DeepL and their final texts obtained a lower HTER. They needed on average approximately half the time for DeepL necessary for MTH. As PE effort is usually lower when the output is of better quality (Kit et Wong, 2015), we can infer from this test that the intrinsic quality of DeepL's output is better than that of our MTH system.

3.3. Comparative evaluation of the post-editing results

This second human evaluation was conducted to ensure that a shorter PE time and a lower HTER do not affect the quality of the final translation. For this evaluation, we asked three Master students in translation (French native speakers) to perform a comparative quality evaluation on the texts that had been post-edited in the post-editing productivity test. We proceeded with a cross-over design, so that the evaluators would not know from which system each segment originated. We asked them to compare the post-edited segments from DeepL and MTH and to indicate which of the translations was the best, or if they considered both segments equivalent in terms of quality. As not all evaluators had German in their language combinations, we provided them with a reference translation. Table 3 presents the number of segments judged as better by a majority of judges (at least 2) for each system, the number of segments judged as equivalent and the number of segments for which no majority emerged. This evaluation gave a Light's kappa score (Light, 1971) of 0.226, that is, according to Lands and Koch's scale (Landis et Koch, 1977), a "fair" agreement.

DeepL better	MTH better	Equivalent	No majority	Total
209 (41.80%)	135 (27.00%)	88 (17.60%)	68 (13.60%)	500 (100%)

Table 3: Number of segments judged as better or equivalent by a majority of judges for each system (in percentage out of the total number of segments).

These results show that for a majority of segments (41.80%), the translation (after PE) originating from DeepL is judged as better than the one originating from MTH. It seems then that a shorter PE time and a lower HTER does not negatively affect the quality of the post-edited translation. When using DeepL, the final output seems to be of better quality for most of the segment.

4. BLEU score's reliability

As our automatic evaluation shows, DeepL performs slightly better than MTH on the test set, the human evaluation, however, gives a clear advantage to DeepL. The translation produced by the neural system was faster to post-edit and required less modification than the one produced by the statistical system and the quality of the final output also tended to be better. This led us to question the reliability of the BLEU score in our context. Two successive studies by Shterionov et al. (2017; 2018) showed that BLEU tends to underestimate the quality of NMT. According to the authors, this underestimation is due to the fact that BLEU, as an n-gram based metric, is better suited for the evaluation of n-gram based systems. Furthermore, NMT tends to produce translations with a length, word order, and word choice that are different from the reference, which tends to lower the BLEU. To verify this hypothesis, we compared the results of our first human evaluation with the BLEU scores at a segment level. We decided to follow a method that is similar to the one used by Shterionov et al. and calculated the *underestimation rate* using the formula introduced by the authors (Shterionov et al., 2017)[3].

We first calculated the BLEU for each of the segments from the corpus used in the PE evaluation. We then counted the segments from DeepL that had a higher "post-editability" (i.e. segments with lower PE time and lower HTER for both evaluators) than their MTH counterparts. Among those segments, we counted the number of segments that had a lower BLEU. We did the same for the segments originating from MTH. To obtain the underestimation rate of BLEU for each system, we divided the number of segments from the system with higher "post-editability" and lower BLEU by the total number of segments with higher "post-editability". Table 4 shows the *underestimation rate* of BLEU for MTH and DeepL.

	Number of segments with higher post-editability	**Number of segments with higher post-editability but lower BLEU**	**% of underestimated segments**
DeepL	144	63	43.75%
MTH	15	5	33.33%

Table 4: Underestimation rate of BLEU for MTH and DeepL.

The *underestimation rate* of BLEU is higher for DeepL (43.75%) than for MTH (33.33%). The results obtained support our hypothesis that BLEU might underestimate the quality of NMT systems.

5. Conclusion

The goal of our study was to determine if a generic NMT system was able to compete with a customised SMT system in our context, i.e., the use of MT as a pre-translation tool at Swiss Post's Language Service. The automatic evaluation based on BLEU indicates that the translation produced by DeepL was slightly better than the one produced by MTH. Our task-based human evaluation clearly indicates that the translation produced by DeepL is better than

[3] The formula suggested by Shterionov et al. is as follows: $\frac{d_{PBSMT}^{NMT}}{d^{NMT}}$ where d^{NMT} is the number of segments from NMT judged as better than their SMT counterparts by human evaluation and d_{PBSMT}^{NMT} the number of segments from d^{NMT} that have a BLEU lower than their SMT counterparts.

the one produced by MTH. The output produced by DeepL was faster to post-edit and required fewer corrections. Furthermore, with the NMT system, the average PE time for the two evaluators is 53.6% lower and the HTER is 75.1% lower. As the PE effort generally reflects the quality of MT, we can then assume that DeepL produces a translation of better quality. This result is corroborated by our second human evaluation assessing the quality of the post-edited translation, which shows that the final translation tends to be of better quality when using DeepL instead of MTH. Regarding the correlation between human and automatic evaluation, we saw that BLEU tends to underestimate the quality of the output of DeepL. This corroborates the hypothesis that BLEU might underestimate the quality of NMT. However, our small-scale study presents some limitations. The use of several automatic metrics might have help us to obtain a more reliable automatic evaluation. Our human evaluations were performed on small corpora and with a limited number of judges, and our results, while relatively clear cut, should be confirmed by a larger-scale study.

Acknowledgements

We would like to thank the evaluators who kindly accepted to participate in our evaluation as well as the Swiss Post and its Language Service.

References

Bouillon, Pierrette, Sabrina Girletti, Paula Estrella, Jonathan Mutal, Martina Bellodi and Beatrice Bircher. 2018. Integrating MT at Swiss Post's Language Service: preliminary results. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation.*, pages 281-286.

Kit, Chunyu and Tak-ming Wong 2015. Evaluation in machine translation and computer-aided translation. In Chan, Sin-Wai (ed.) *The Routledge encyclopedia of translation technology.* London : Routledge, pages 213-236.

Landis, Richard J. and Gary G Koch 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics,* vol. 33 (1)**,** pages 159-174.

Light, Richard J. 1971. Measures of response agreement for qualitative date: some generalization and alternatives. *Psychological Bulletin,* vol. 76 (5)**,** pages 365-377.

Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Associationfor Computational Linguistics.* Association for Computational Linguistics, pages 311-318.

Shterionov, Dimitar, Pat Nagle, Laura Casanellas, Riccardo Superbo and O'Dowdm Tony. 2017. Empirical evaluation of NMT and PBSMT quality for large-scale translation production. In *Proceedings of the the 20th Annual Conference of the European Association for Machine Translation.* pages 74-79.

Shterionov, Dimitar, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'Dowd and Andy Way 2018. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation,* vol.**,** pages 1-19.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Micciulla Linnea and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas.* (Cambridge) The Association for Machine Translation in the Americas, pages 223-231.

TAUS and CNGL 2010. *MT Post-editing Guidelines*. URL: https://www.taus.net/academy/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines [last accessed 02 March 2018].

Volkart, Lise. 2018. Traduction automatique statistique vs. neuronale : comparaison de MTH et DeepL à la Poste Suisse. Master, Geneva.