Differences between SMT and NMT Output - a Translators' Point of View

Jonathan Mutal[1], **Lise Volkart**[1], **Pierrette Bouillon**[1], **Sabrina Girletti**[1], and **Paula Estrella**[2]

[1]FTI/TIM, University of Geneva, Switzerland
[2]FaMaF y FL, University of Córdoba, Argentina
{Jonathan.Mutal,Lise.Volkart}@unige.ch
{Pierrette.Bouillon,Sabrina.Girletti}@unige.ch
paula.estrella@unc.edu.ar

Abstract

In this study, we compare the output quality of two MT systems, a statistical (SMT) and a neural (NMT) engine, customised for Swiss Post's Language Service using the same training data. We focus on the point of view of professional translators and investigate how they perceive the differences between the MT output and a human reference (namely deletions, substitutions, insertions and word order). Our findings show that translators more frequently consider these differences to be errors in SMT than NMT, and that deletions are the most serious errors in both architectures. We also observe there to be less agreement on differences to be corrected in NMT than SMT, suggesting that errors are easier to identify in SMT. These findings confirm the ability of NMT to produce correct paraphrases, which could also explain why BLEU is often considered to be an inadequate metric to evaluate the performance of NMT systems.

1 Introduction

Some recent studies have investigated the differences between statistical machine translation (SMT) and neural machine translation (NMT) in terms of the quality of the output (Daems and Macken, 2019; Toral and Cartagena, 2017; Bentivogli et al., 2016). In this paper, we focus on the point of view of professional translators and investigate how they perceive the differences in the translations produced by a SMT and a NMT system, both trained on the same data for comparison purposes.

Since we cannot evaluate all the differences, we will only look at divergent cases, that is, where one type of system (SMT or NMT) produces a sentence which is identical or very close to a human reference translation, while the other produces a different translation. We want to answer the following research questions: 1) What are the differences between SMT and NMT in terms of *edits* needed to reach the Post's official reference (namely deletions, substitutions, insertions and word order)?, 2) Would translators post-edit these differences? and, finally, 3) Do the translators agree on this task? Our hypothesis is that the type of edits differs between NMT and SMT and that with NMT, edits will be less often considered as real errors by translators.

In the following sections, we will describe the context of this study, the test data and how we built the SMT and NMT engines. We will then describe the methodology used for the evaluation and the results obtained.

2 Context, MT Engine Training and Test Data

This study is part of a collaboration between the University of Geneva and Swiss Post's in-house Language Service (Bouillon et al., 2018). The Language Service translates a broad range of texts from and into German, French, Italian and English. In the context of testing two MT architectures (SMT and NMT), we are interested in discovering which differences between the MT output and the reference translation are considered by the translators to be errors worth editing.

Our analysis focuses on two customised machine translation engines for the language pair German-to-French, a neural and a statistical one, trained with the same training data. The train-

Temnikova, I, Orăsan, C., Corpas Pastor, G., and Mitkov, R. (eds.) (2019) *Proceedings of the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, Varna, Bulgaria, September 5 - 6, pages 75–81.
https://doi.org/10.26615/issn.2683-0078.2019_009

ing data consisted of 2,558,148 translation units from the main translation memory of Swiss Post's Language Service. In order to avoid dealing with different variables that interfere with the real objective of this evaluation, such as pre-processing, post-processing and tune hyper-parameters, we kept the training as simple as possible for both architectures.

SMT engine. We followed the training process (corpus tokenization, language and translation model training, tuning and testing on a disjoint set from training) using the tools provided by Moses[1]. Language models were trained using KenLM (Heafield, 2011) on 4-grams.

NMT engine. We segmented infrequent words into their corresponding sub-word units by applying the byte pair encoding (BPE) approach (Sennrich et al., 2015); an encoder-decoder NMT model, transformer (Vaswani et al., 2017), was then trained using OpenNMT-tf (Klein et al., 2017). For this model, we used default hyper-parameters[2].

Subset	#sentences	#tokens	#vocabulary
Train	2M	36M	618k
Dev	100k	1.6M	112k
Test	1k	23k	4k

Table 1: Number of sentences, tokens and vocabulary for German (source language).

Subset	#sentences	#tokens	#vocabulary
Train	2M	40M	252k
Dev	100k	2.1M	56k
Test	1k	32k	3k

Table 2: Number of sentences, tokens and vocabulary for French (target language).

Test data. In order to evaluate both models, we built a development data set by extracting 5% of the sentence pairs from the training data. The test data consist of 1,736 translation units retrieved from process manuals. Tables 1 and 2 summarise the number of sentences, tokens and vocabulary for each subset in each language.

3 Methodology

In order to compare the two architectures and answer our research questions, we performed both an automatic and human evaluation with professional translators from Swiss Post's Language Service. In the literature, many error taxonomies have been used to carry out MT evaluations (Daems et al., 2017; Lommel et al., 2014; Stymne and Ahrenberg, 2012). In this study, we focus instead on type of edits, namely (*i*) word insertions, (*ii*) word deletions, (*iii*) word substitutions, and (*iv*) word order.

3.1 Automatic Evaluation

Two standard MT metrics were used to measure the performance of both architectures on the complete test set: TER (Snover et al., 2006) and BLEU (Papineni et al., 2002). The different types of edits (substitutions, deletions, word order and insertions) were also automatically calculated using TER.

3.2 Human Evaluation

In order to compare the two types of systems (SMT and NMT), we decided to focus on translations that are different in the two architectures and are close to the reference from the translation memory (see Section 2) in one architecture, but more distant in the other. These sentences are interesting since at least one of the systems was able to produce a good translation.

We selected the two sets of data using BLEU. The first (SMT-div) contains all sentences for which NMT obtains a high BLEU score ($>$ 85) and SMT a lower score ($<$ 85) (353 sentences). The second (NMT-div) includes sentences with a high BLEU score in SMT ($>$ 85) and a lower one in NMT ($<$ 85) (77 sentences).

For this human evaluation, we decided to manually identify the edits (insertions, substitutions, etc.) in order to group successive edits in one single edit, for example the two insertions (”sont autorisés”) and the substitution (”peuvent” by ”à”) were grouped in an single substitution ” sont autorisés à”, as illustrated in Table 3. In that way, we identified 143 edits in the test set NMT-div and 675 in the SMT-div. As we were conducting a qualitative study and due to time constraint for the human evaluation, we decided to evaluate the same number of edits for both systems. We randomly extracted 143 edits from SMT-div to build the final test sets. In each test set, the edits were

[1] For training processes, see: http://www.statmt.org/moses/?n=Moses.Baseline

[2] http://opennmt.net/OpenNMT-tf/model.html#catalog

TER	MT output	Human annotation	Type
Substitution Deletion (forces de)	evénements dus aux éléments naturels (tremblements de de terre, inondations, etc.)	evénements dus aux éléments naturels (tremblements de de terre, inondations, etc.)	Substitution
Reference: *événements dus aux forces de la nature (tremblement de terre, inondation, etc.)*			
Insertion Substitution	les filiales sont autorisées à vérifier certains groupes de marchandises plus souvent. les filiales sont autorisées à vérifier certains groupes de marchandises plus souvent.	les filiales sont autorisées à vérifier certains groupes de marchandises plus souvent.	Substitution
Reference: *les filiales peuvent vérifier certains groupes de marchandises plus souvent.*			

Table 3: Examples of grouping multiple edits into a single edit.

Source	MT Output	Edits
der Abholer ist persönlich bekannt:	la personne qui vient retirer l'envoi est connue personnellement:	Insertion
Reference: *cette personne est connue personnellement:*		
immer die Adresse der Filiale aufführen, nicht diejenige des Hauptsitzes.	toujours indiquer l'adresse de la filiale, et non celle du siège principal.	Substitution
Reference: *toujours mentionner l'adresse de la filiale, et non celle du siège principal.*		
mit einer Zustellliste XXX werden mehrere Sendungen auf einer Liste zusammengeführt.	plusieurs envois sont regroupés sur une liste avec une feuille de distribution XXX.	Word order
Reference: *avec une feuille de distribution XXX, plusieurs envois sont regroupés sur une liste.*		

Table 4: Examples of sentences with edits in colour

highlighted in red. In order to evaluate the edits individually, we duplicated the sentences containing more than one edit, and we marked only one edit at a time. Three translators from Swiss Post's Language Service received these target sentences in a spreadsheet along with the source sentences. For each edit, they had to state if they would modify the red part during a full post-editing task. They were not asked to post-edit the sentences, but only to indicate if they would change the highlighted part or not. Table 4 shows three different sentences with edits marked in red (as presented to the evaluators), as well as the corresponding reference translations. During the evaluation task, the evaluators did not have access to the reference translation and had no information about the type of system used to produced the output.

Results were collected. We calculated 1) how many differences post-editors would change in both systems, 2) the corresponding type of edit and 3) the inter-rater agreement.

4 Results

4.1 Automatic Evaluation

The two systems obtained high BLEU scores on the test set (1,736 sentences), 0.68 for NMT and 0.59 for SMT, and low TER scores of 19.96 and 30.05, showing that both systems produce good quality translations according to automatic evaluation.

Table 5 shows the number of substitutions, insertions, deletions and word order differences in both architectures. The total number of edits is higher for SMT than NMT, with a total of 10,399 and 7,327 edits respectively.

For both systems, the most frequent type of edits are substitutions, followed by deletions, insertions and word order. However, the proportion of deletions is higher for SMT than NMT (36% vs 27%), whereas the proportion of substitutions is higher for NMT (47% vs 37%).

Table 6 shows the number of edits in the output

Edit	**SMT**	**NMT**
Insertions	1,869 (18%)	1,305 (18%)
Deletions	3,754 (36%)	1,995 (27%)
Substitutions	3,881 (37%)	3,470 (47%)
Word order	895 (9%)	557 (8%)
Total	10,399 (100%)	7,327 (100%)

Table 5: Number of edits and percentage per edit in SMT vs NMT for language pair German-to-French.

sentences for items where SMT obtained a higher BLEU score than NMT (396 sentences). Table 7 shows the number of edits in the reverse situation (1,003 sentences). For the 424 remaining sentences, the translations by both systems obtained identical BLEU scores (100 BLEU point).

Edit	**SMT**	**NMT**
Insertions	342	547
Deletions	758	820
Substitutions	670	1333
Word order	144	252

Table 6: Number of edits in sentences where SMT has a higher BLEU score than NMT (396 sentences).

Edit	**SMT**	**NMT**
Insertions	1461	690
Deletions	2911	1092
Substitutions	3044	1837
Word order	1467	289

Table 7: Number of edits in sentence output where NMT has a higher BLEU score than SMT (1003 sentences).

It can be observed that when SMT has a higher BLEU, NMT almost doubles the number of substitutions (by 1.95) and word order (by 1.75) compared to SMT, whereas when NMT is better, all types of edits double, with word order edits being multiplied by 5.07. This means that when NMT is good, SMT produces more word order difference, as shown in example (Table 8).

Overall, the most common edit is substitution for both systems. However, if we compare the percentage of edits in both architectures, the number of substitutions is much higher in NMT (47.34%), which can be explained by the well-known ability of NMT to paraphrase (Mallinson et al., 2017). We can see a clear example in Table 9. On the other hand, SMT had more deletions (36.09%). For the other types of edits, there is not much difference between the two systems.

4.2 Human Evaluation

The aim of the human evaluation is to shed light on how translators perceive edits in the output of each system, namely whether they would edit them or not. We also wanted to determine which types of edits would be post-edited more often by translators.

For each sentence, we considered the majority judgement (at least 2 judges agree) and we computed the results for both test subsets (143 edits per system). Figure 1 shows the percentages of edits that a majority of judges would change, per system and per type of edit.

If we consider all edits together, the evaluators would have post-edited the SMT output more than the NMT output: 68.53% of the edits would have been modified by a majority of judges in SMT versus 14.69% in NMT. This confirms our hypothesis that the edits in NMT are more often considered to be non-significant in the post-editing task.

For both systems, the edit type most frequently marked by the translators as something they would modify was deletions, which is not surprising since an omission in the output will very likely affect the quality of the translation. As for substitutions, which was the most frequent edit in both systems (see Tables 5 and 10), the majority of judges would modify more than half of them (62.82%) in SMT output vs only 14.81% in NMT. This illustrates the ability of neural systems to paraphrase and use correct synonyms. Finally, we can see that word order differences, which increase in SMT when NMT is better (Table 7), were mostly considered to be mistakes in SMT, which reflects the well-known fact that SMT has problems dealing with word order differences.

We also looked at the agreement between judges on this task. We computed Light's Kappa (Light, 1971) for the SMT and NMT evaluation. For SMT overall, we obtained a Kappa of 0.332 with a high statistical significance of evidence (p-value of 0.6%), corresponding to a fair agreement. For NMT overall, however, we obtained a Kappa of 0.166 which represents a slight agreement (Landis and Koch, 1977), but with a low statistical

Source	**SMT output**
suchen Sie die Räumlichkeiten und die Umgebung der Filiale bis zum Eintreffen der Polizei nach verdächtigen Gegenständen ab.	fouillez les locaux et les environs de la filiale jusqu'à l'arrivée de la police après d'objets suspects.
Reference: *fouillez les locaux et les environs de la filiale à la recherche d'objets suspects jusqu'à l'arrivée de la police.*	

Table 8: An example of word order error for SMT.

Source	**NMT output**
der zuständige Geschäftsbereich übernimmt die interne Information und leitet bei Bedarf Massnahmen ein.	l'unité d'affaires compétente prend en charge l'information interne et prend des mesures si nécessaire.
Reference: *l'unité d'affaires compétente assure l'information interne et met en œuvre des mesures en cas de besoin.*	

Table 9: An example of substitution for NMT.

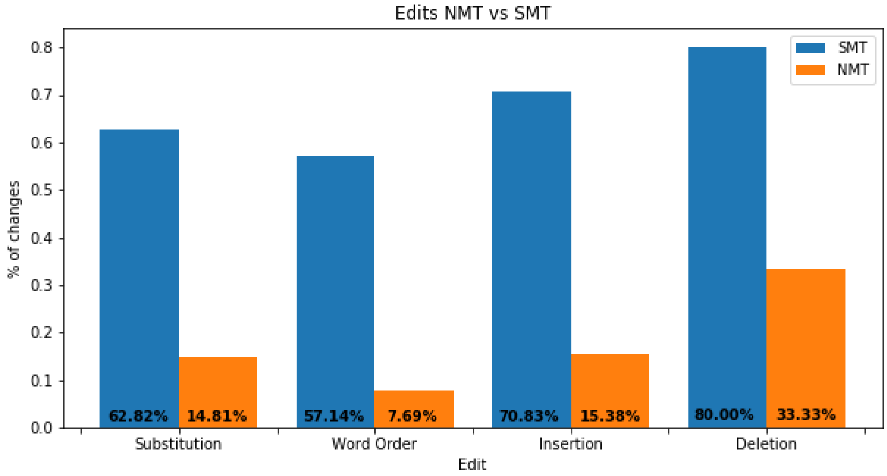

Figure 1: % of edits the translators would modify for SMT and NMT (by at least two judges).

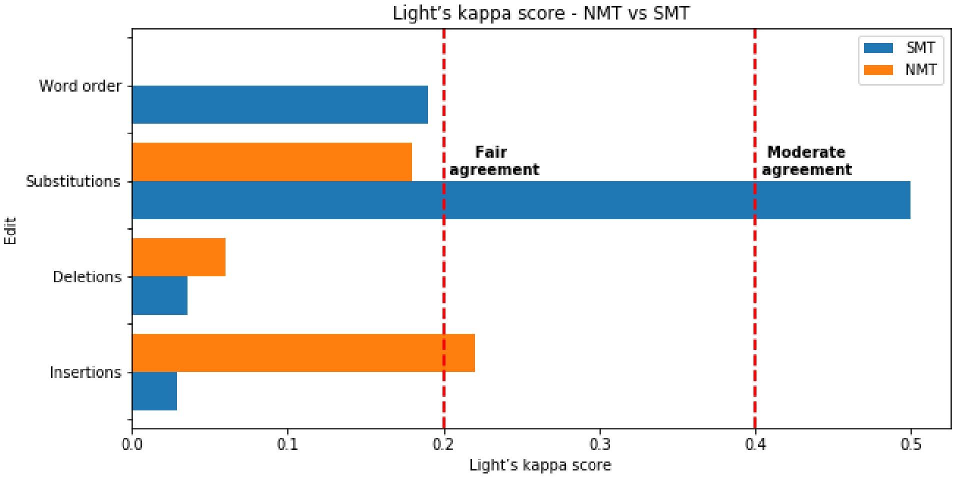

Figure 2: Agreement for each type of edit.

Edit	**SMT**	**NMT**
Insertions	16.78%	10%
Deletions	17.48%	4.19%
Substitutions	54.54%	75.52%
Word order	11.18%	10%

Table 10: %edit type in 143 edits extracted from each model.

significance (p-value 34%).

Figure 2 illustrates individual Light's kappa scores computed for each edit type. These scores show that judges do not strongly agree on the divergences that would need post-editing, particularly with NMT output. In particular, evaluators disagree on the word order category for NMT output, where the Light's kappa score obtained is negative. Translators moderately agreed (K=0.50) on substitutions in SMT (p-value<0.0011) and fairly on insertions (K=0.22) in NMT (p-value>0.62) (see Figure 1). This suggests that in NMT, translators have more difficulties clearly stating whether a sentence has to be modified or not.

5 Conclusion and Future Work

In this paper, we presented an innovative methodology to compare SMT and NMT based on differences with an official reference. We showed that (*i*) the most common edits are substitutions, with respectively 37.32% and 47.34% for NMT and SMT, and deletions with 27% and 36.09%; (*ii*) the most significant difference from a translator's point of view is deletions, in particular in SMT, with 80% of changes in SMT but only 33.33% in NMT; (*iii*) NMT edits are more often considered to be non-significant from a post-editing point of view (14.68%), as opposed to SMT edits (68.53%); (*iv*) translators have more difficulties stating whether a sentence has to be modified with NMT than with SMT.

This study has several limitations: three judges were not enough to obtain a good inter-agreement score. It will be interesting to test the same methodology with the different languages of the Post (Italian and English) in order to see if there are cross-lingual differences, as well as with translators trained for post-editing. We also would like to see if differences considered to be wrong by translators are related to specific types of errors.

However, despite its limitations, the paper provides interesting perspectives. Firstly, the fact that NMT produces correct paraphrases of the reference confirms a common hypothesis that BLEU is not an adequate metric for evaluating the performance of NMT (Shterionov et al., 2017, 2018; Volkart et al., 2018). From a broader perspective, the collected data, which focus on different types of individual edits, could also be used to train translators on how to distinguish between essential vs non essential changes.

Acknowledgement

We would like to thank the translators who kindly accepted to participate in this study, as well as Martina Bellodi and the whole Swiss Post's Language Service.

References

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 257–267. https://doi.org/10.18653/v1/D16-1025.

Pierrette Bouillon, Paula Estrella, Sabrina Girletti, Jonathan Mutal, Martina Bellodi, and Beatrice Bircher. 2018. *Integrating MT at Swiss Post's Language Service: preliminary results*, pages 281–286. Proceedings of the 21st Annual Conference of the European Association for Machine Translation. ID: unige:105252. https://archive-ouverte.unige.ch/unige:105252.

Joke Daems and Lieve Macken. 2019. Interactive adaptive smt versus interactive adaptive nmt: a user experience evaluation. *Machine Translation* 33(1):117–134. https://doi.org/10.1007/s10590-019-09230-z.

Joke Daems, Sonia Vandepitte, Robert Hartsuiker, and Lieve Macken. 2017. Identifying the machine translation error types with the greatest impact on post-editing effort. *Frontiers in Psychology* 8:1282. https://doi.org/10.3389/fpsyg.2017.01282.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Edinburgh, Scotland, pages 187–197. https://www.aclweb.org/anthology/W11-2123.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1).

Richard Light. 1971. Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin* 76:365–377.

Arle Lommel, Aljoscha Burchardt, Maja Popovic, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of mt errors on real data. pages 165–172.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, pages 881–893. https://www.aclweb.org/anthology/E17-1083.

Kishore Papineni, 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 on Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '02, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR* abs/1508.07909. http://arxiv.org/abs/1508.07909.

Dimitar Shterionov, Pat Nagle, Laura Casanellas, Riccardo Superbo, and Tony O'Dowd. 2017. Empirical evaluation of nmt and pbsmt quality for large-scale translation production.

Dimitar Shterionov, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'dowd, and Andy Way. 2018. Human versus automatic quality evaluation of nmt and pbsmt. *Machine Translation* 32(3):217–235. https://doi.org/10.1007/s10590-018-9220-z.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*. pages 223–231.

Sara Stymne and Lars Ahrenberg. 2012. On the practice of error analysis for machine translation evaluation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. European Languages Resources Association (ELRA), Istanbul, Turkey, pages 1785–1790. http://www.lrec-conf.org/proceedings/lrec2012/pdf/717$_{paper.pdf}$.

Antonio Toral and Víctor Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. pages 1063–1073. https://doi.org/10.18653/v1/E17-1100.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR* abs/1706.03762. http://arxiv.org/abs/1706.03762.

Lise Volkart, Pierrette Bouillon, and Sabrina Girletti. 2018. *Statistical vs. Neural Machine Translation: A Comparison of MTH and DeepL at Swiss Post's Language Service*, pages 145–150. Proceedings of the 40th Conference Translating and the Computer. ID: unige:111777. https://archive-ouverte.unige.ch/unige:111777.