Detecting Post-edited References and Their Effect on Human Evaluation

Věra Kloudová, Ondřej Bojar, Martin Popel
Charles University,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics,
Malostranské náměstí 25, 118 00
Prague, Czech Republic
{bojar,kloudova,popel}@ufal.mff.cuni.cz

Abstract

This paper provides a quick overview of possible methods how to detect that reference translations were actually created by post-editing an MT system. Two methods based on automatic metrics are presented: BLEU difference between the suspected MT and some other good MT and BLEU difference using additional references. These two methods revealed a suspicion that the WMT 2020 Czech reference is based on MT. The suspicion was confirmed in a manual analysis by finding concrete proofs of the post-editing procedure in particular sentences. Finally, a typology of post-editing changes is presented where typical errors or changes made by the post-editor or errors adopted from the MT are classified.

1 Introduction

Over ten years of WMT (Conference on Machine Translation, Barrault et al., 2020)[1] saw a number of manual evaluation methods and established the best strategies for obtaining reference translations for automatic evaluation, see Appendix B in WMT 2020 Findings (Barrault et al., 2020).

One of the instructions for preparing the reference translations explicitly prohibits using any machine translation. Yet, in 2020, one of the agencies has not followed this instruction. Not only was it easy to recognize, but we learned several novel insights into manual evaluation of translation, by examining post-edited and independent reference translations and providing a small contrastive style of manual evaluation.

2 Dataset

We used the English-Czech part of WMT2020 (Barrault et al., 2020) news test set, which consists of 130 documents (1418 segments) originally written in English – news stories downloaded from web. The test set comes with an official reference translation into Czech (REF1) provided by the WMT organizers and done by a professional translation agency. There are also 8 machine translations submitted by the participants of the WMT news translation shared task and 4 translations by online systems anonymized as ONLINE-A, ONLINE-B, ONLINE-G and ONLINE-Z.

We focused on three translations: the official reference, REF1; the best-performing MT system (according to the official WMT manual evaluation), CUNI-DOCTRANSFORMER (Popel, 2020); and the best-performing online system, ONLINE-B.

We hired two professional translators (native Czech speakers) to translate the whole WMT20 test set, thus creating additional references REF2 and REF3. We also hired 18 annotators to judge the translation quality of REF1, CUNI-DOCTRANSFORMER and ONLINE-B.[2] The annotators assessed 90 of the 130 documents, using the RankME evaluation (Novikova et al., 2018) following the methodology of Popel et al. (2020). In this RankME evaluation, fluency, adequacy and overall quality are evaluated in a source-based sentence-level document-aware fashion, on a 0–10 scale, where all the evaluated translations are shown on the same screen, allowing thus better reliability in comparisons; see Section 5 for details.

3 Automatic analysis of references

Table 1 shows the translation quality of the three references and two selected MT systems according to two manual evaluations, DA (Direct Assessment, Graham et al., 2013) and RankME, and four types of BLEU scores. The first three types use

[1] http://www.statmt.org/wmt06 till wmt20

[2] The additional references REF2 and REF3 were not available before our RankME evaluation started. We plan to evaluate them in future.

Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval), pages 114–119
Online, April 19, 2021.

system	manual		BLEU			
	DA	RankME	REF1	REF2	REF3	REF2+3
REF1	**85.6**	**8.17**	–	28.91	24.18	37.22
REF2	–	–	28.90	–	26.43	–
REF3	–	–	24.20	26.45	–	–
CUNI-DOCTRANSFORMER	82.8	7.39	35.88	**36.50**	**30.17**	**47.59**
ONLINE-B	70.5	5.62	**41.11**	31.08	26.39	41.00

Table 1: Manual and automatic evaluation scores of the systems in our study. DA is the source-based Direct Assessment average score (un-normalized). RankME is the average Overall quality score over all 90 documents (not sentences) evaluated in our study. BLEU is computed with SacreBLEU (Post, 2018) with signature BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.4.13. The best score in each column is in bold.

REF1, REF2 and REF3, respectively, as the reference translation in BLEU. The fourth type uses BLEU with two reference translations: REF2+3.

While both manual evaluations, DA and RankME, agree that REF1 is better than both CUNI-DOCTRANSFORMER and ONLINE-B, the automatic metric BLEU evaluates one of the two MT systems as better than REF1.[3] For brevity, we report only BLEU, but we confirmed this with several other automatic metrics, e.g. chrF (Popović, 2015).

The reason for this surprising observation is that most sentences in REF1 are actually post-edited versions of ONLINE-B, as we show in Sections 3.1 and 4 and as was acknowledged by the agency after our investigation. Thus, REF1 and ONLINE-B are more similar than if REF1 had been translated from scratch.[4]

It is well-known that BLEU (and other automatic metrics based on similarity with reference translations) is biased when evaluating a system which was used as a basis for post-editing the reference translation.

It is important to note that the official (manual) evaluation carried out by WMT for the affected English-to-Czech translation direction was *source-based* DA. The annotators of the DA thus were not affected by the quality of references; instead, they blindly rated REF1 as if it was another competing translation system. The resulting scores of DA document that source-level DA is sufficiently reliable, robust to invalid references. At the same time, it was a little surprising to us that "mere post-editing" can increase translation quality so substantially that REF1 significantly outperformed all other systems. Despite the remaining translation errors in REF1, see below, the translator/post-editor did the job well.

For a finer analysis, we process the news test set at the level of individual documents in the following.

3.1 Automatic detection of post-editing

Our first suspicion that REF1 is actually post-edited ONLINE-B stems from the fact that ONLINE-B achieved the highest $\text{BLEU}_{\text{REF1}}$ score (i.e. BLEU with REF1 as the reference) out of all the MT systems, including the best one according to the manual evaluation, CUNI-DOCTRANSFORMER, as shown in Table 1. In order to confirm this suspicion, we wanted to automatically find documents where the probability of being post-edited is the highest.

Below, we suggest two methods for such document-level automatic detection of post-editing. The first method needs just an output of another MT system. The second method needs one or more additional human references.

3.1.1 Detection using another MT system

We selected CUNI-DOCTRANSFORMER as the other MT system for two reasons. First, it is the best MT system in English-Czech WMT20 according to the official manual evaluation. Second, as far

[3] Actually, both MT systems are better than all three references according to all BLEU scores, with a single exception of $\text{BLEU}_{\text{REF3}}$(ONLINE-B) = 26.39 < 26.43 = $\text{BLEU}_{\text{REF3}}$(REF2), which is not statistically significant (bootstrap resampling, $p < 0.05$). Obviously, we cannot use e.g. $\text{BLEU}_{\text{REF1}}$ to judge the quality of REF1 (it would be 100, by definition).

[4] One of the instructions for the translation agency preparing references for WMT 2020 was: *All translations should be "from scratch", without post-editing from MT. Using post-editing would bias the evaluation, so we need to avoid it. We can detect post-editing so will reject translations that are post-edited.* (Barrault et al., 2020). Unfortunately, the WMT organizers did not detect the post-editing in this case (as we do in this paper) and did not reject the translation.

as we know, CUNI-DocTransformer was not available online at the time of creating the WMT20 references, so it could not be used as the basis for post-editing Ref1.

For each document d, we computed

$$Detection1 = \text{BLEU}_{\text{Ref1}}(\text{Online-B}, d) - \text{BLEU}_{\text{Ref1}}(\text{CUNI-DocTransformer}, d). \quad (1)$$

This score was positive for 104 out of the 130 documents. In other words, for 80% of documents, the reference is more similar to Online-B than to the best-performing CUNI-DocTransformer, a likely indicator that most of the documents were post-edited.

We inspected manually three documents with the most negative $Detection1$ score and did not find any clues of post-editing, but we noticed the quality of Online-B was low for these documents, so perhaps the translator throw away the MT output and translated these documents from scratch (or post-edited so heavily that the original MT output cannot be detected).

We inspected manually three documents with the most positive $Detection1$ score and observed these well translated with a reasonably high quality by Online-B, and required just few minor post-edits, as was done in Ref1.

Finally, we inspected sentences from other documents and found further signals of post-editing (even when the $Detection1$ score was not high enough to be a convincing proof alone). These examples are discussed in Section 4.

3.1.2 Detection using additional references

A similar detection can be used with an additional human reference instead of an MT system. We had available two such references, Ref2 and Ref3. We thus opted for a slightly different detection formula which allows to use two-reference BLEU:

For each document d, we computed

$$Detection2 = \text{BLEU}_{\text{Online-B}}(\text{Ref1}, d) - \text{BLEU}_{\text{Ref2+3}}(\text{Ref1}, d). \quad (2)$$

This resulted in a similar ordering of documents as the $Detection1$ method.

3.2 Human translation is more similar to MT than other humans

For each document d, we computed

$$Score3 = \text{BLEU}_{\text{Ref2+3}}(\text{Ref1}, d) - \text{BLEU}_{\text{Ref2+3}}(\text{Online-B}, d). \quad (3)$$

This $Score3$ score was negative for 96 out of the 130 documents. This means that the similarity of Online-B to the additional references Ref2 and Ref3 is higher than the similarity of Ref1 to Ref2 and Ref3. When focusing just on Ref2, we can see that $\text{BLEU}_{\text{Ref2}}(\text{OnlB}) = 31.08 > 28.91 = \text{BLEU}_{\text{Ref2}}(\text{Ref1})$. This is very surprising given our hypothesis that Ref2 is actually translated from scratch without any post-editing.

We have two possible explanations for this. First, the Ref1 translator tried to "hide" the fact that the translation is post-edited, by doing edits which do not affect the translation quality. Second, the Ref1 translator actually improved the Online-B translation quality by post-edits which result in less literal translations, while the Ref2 translator opted more frequently translations which were likely to be independently produced also by the MT system.

Given the fact that both DA and RankME manual evaluations show Ref1 is significantly better than Online-B, we hypothesize most of the post-edits were actually improvements. We noticed just a few opposite cases (see below, category 3).

4 Post-editing changes typology

In this part of our study, we would like to present a classification of post-editing changes observed in texts which we claim to be post-edited machine translations. These changes signal that the reference translation Ref1 has been actually created by post-editing a MT system. For this purpose, we used MT-ComparEval (Klejch et al., 2015) to select 27 sentences which show the highest n-gram overlap with the suspected MT system. We analyzed the edits made by a post-editor in a MT output and compared the source text (English), the MT output (Czech) and the output of the manual post-editing process (Czech).

Based on the particular changes found after comparing these three versions of our sentences, we defined the following categories (with examples where Src is the source sentence, OnlB is the MT Online-B and Ref1 is the human post-editing). In each category, we would like to present particularly noticeable changes which we assume to be clear evidence of the post-editing procedure. We classify these changes into three categories:

- Category 1: minor changes, particularly focused on grammar categories, in long adopted structures from the MT (preserving or improving the overall quality of the output),

- Category 2: unnecessary shifts (without a significant impact on the quality of the output),
- Category 3: negative shifts (errors made by the post-editor, preserving or even worsening the quality of the MT output).

Furthermore, we also found these errors or conspicuous structures in the post-edited output:

- Category 4: errors adopted from the MT which the post-editor has not discovered; therefore, they have been preserved in the final text of REF1.

For all these four categories, we can state they prove the final output is a result of a post-editing process.

4.1 Typology 1: Changes by the post-editor

4.1.1 Changes in spelling

Category 3: Errors in writing:

SRC their 17-month-old daughter
ONLB jejich **17měsíční** dcerou
REF1 jejich **17timěsíční** dcerou

The standard Czech grammar allows only forms *17měsíční* or *sedmnáctiměsíční* (17-month). The form *17timěsíční* (where "*ti*" reflects the pronunciation) is considered an error.

4.1.2 Grammatical changes

Category 1: Changes in verb tenses:

SRC Wang's death follows after those of other activists [. . .]
ONLB Wangova smrt **následuje** po těch dalších aktivistech [. . .]
REF1 Wangova smrt **následovala** po úmrtí dalších aktivistů [. . .]

The present tense (*následuje* = *follows*) used in ONLB was changed into past tense (*následovala* = *followed*). Changes between past and present tense (while keeping the same verb) occurred relatively often (8 cases) in the investigated 27 sentences.

Category 3: Case and noun number changes:

SRC Nobel laureate winner Liu Xiaobo [. . .]
ONLB Nositel Nobelovy ceny **za laureát** Liu Xiaobo [. . .]
REF1 Nositel Nobelovy ceny **laureátů** Liu Xiaobo [. . .]

Despite the changes made, the semantic defectiveness of ONLB has been preserved. In Czech, the Nobel prize is *Nobelova cena* and the word *laureátů* (genitive sg.) is wrong. In this case, *Laureát Nobelovy ceny* would be acceptable.

Category 3: Changes in adjective comparison:

SRC The market was more receptive to lesser-known names [. . .]
ONLB Trh byl **vnímavější** k méně známým jménům [. . .]
REF1 Trh byl **více vnímavý** k méně známým firmám [. . .]

The correct comparative of *vnímavý* = *receptive* is *vnímavější*. The phrase *více vnímavý* is considered non-standard and rather rare.[5]

4.1.3 Changes in accuracy of information

Category 3: Changes incompatible with the meaning of the source text:

SRC from some 3 billion cu ft at the start of 2019
ONLB z **přibližně** 3 miliard cu ft na začátku roku 2019
REF1 ze **současných** 3 miliard kubických stop na začátku roku 2019

ONLB uses *přibližně* = *approximately*, but REF1 changed it to *současných* = *current*, which is wrong because the article was written in September 2019, when the amount was already much higher than 3 billion cu ft, according to the source text.

4.2 Typology 2: Errors adopted from ONLB

4.2.1 Errors in spelling

Category 4: English spelling in foreign names:

SRC [. . .], Uighur scholar
ONLB [. . .], **Uighurský** učenec
REF1 [. . .], **Uighurský** učenec

The correct Czech spelling would be *ujgurský* (with lowercase *u*, in this sentence).

Category 4: MT spelling errors:

SRC Endeavor Group Holdings
ONLB Společnost **Endeavour** Group Holdings
REF1 Společnost **Endeavour** Group Holdings

ONLB introduced a spelling error – *Endeavo**u**r*, which remained unnoticed by REF1.

[5] Such analytic comparatives can be seen as a proof of translationese (Toury, 1995), i.e. in this case, influenced by the English analytic comparative *more receptive*.

4.2.2 Lexical and stylistic errors

Category 4: Errors in lexical meaning:

SRC [...] then evidence showed the officers had reason to believe their lives were in danger [...]

ONLB [...] důkazy ukazují, že **důstojníci** měli důvod se domnívat, že jejich životy jsou v nebezpečí [...]

REF1 [...] podle svědectví důkazů měli **důstojníci** důvod domnívat se, že jejich životy byly v ohrožení [...]

Given the context of the source article, *officers* should be translated as *policisté = police officers*. It is questionable whether the translation *důstojníci = commissioned officers* is acceptable.

Category 4: Errors in meaning of a syntactic structure:

SRC an invite from Khloe to a 'Taco Tuesday' dinner at her mansion

ONLB pozvání od Khloe **na večeři "Taco Tuesday" u jejího sídla**

REF1 pozván od Khloé **na večeři v "Taco Tuesday" u jejího sídla**

In the highlighted example, REF1 added only the preposition *v = in*, which bears out the superficial reading of the machine translation output *a dinner in a "Taco Tuesday" restaurant near her mansion*. However, the original meaning is quite different: "Taco Tuesday" is a custom of going out to eat tacos, not a name of a restaurant.

Category 4: Improper collocations:

SRC The California attorney general's office in March declined to issue state criminal charges after a nearly yearlong investigation.

ONLB Kancelář generálního prokurátora v Kalifornii v březnu po téměř celoročním vyšetřování odmítla **vydat státní trestní obvinění**.

REF1 Kancelář generálního prokurátora v Kalifornii v březnu po téměř celoročním vyšetřování odmítla **podat trestní obvinění**.

In this example, solely the verb prefix changed (*vydat → podat*), but not the noun (*obvinění = accusation, charge*). The correct translation would be *podat trestní oznámení*.

5 Human evaluation – RankME

In our blind RankME evaluation by 18 human judges (6 professional translators, 6 students from MA Study Program Translation: Czech and English at the Institute of Translation Studies, Charles University's Faculty of Arts,[6] and 6 non-professionals with excellent knowledge of the English language), 90 documents (887 segments, typically sentences) were evaluated on a sentence level in terms of adequacy, fluency and overall quality (as defined by Popel et al. (2020)). Every document was scored by two evaluators, and every evaluator scored ten different documents. Then we could compare the ratings of the post-edited REF1 and of the suspected ONLB. According to the ratings, the translation quality is, in all cases, better in REF1 compared to ONLB. The post-editor improved the quality of the ONLB in all three categories: in adequacy (increased by 2.23 on average), fluency (2.70) and overall quality (2.55), on a 0–10 scale. As could be expected, the most significant improvement occurred in fluency. Apparently, the post-editor was more sensitive to errors in fluency rather than adequacy, as shown also in our analysis of post-editing changes.

6 Conclusion

We suspected that WMT20 reference translations were actually post-edits of one of the participating systems, ONLINE-B, and not created independently. We proposed two methods to detect this situation, both confirming our suspicion.

In a subsequent manual analysis, we provided numerous examples of translation choices in the reference translation which are extremely unlikely to happen when translating from scratch.[7]

The result of this analysis is a draft typology of post-editing strategies. We see this typology as an interesting basis for further inspection of translations in a world where post-editing becomes the industry standard.

By contrasting Direct Assessment scores with our manual evaluation (RankME), we observed the post-editor improved primarily *fluency* of the translation and less so its adequacy. It would be useful to confirm this observation on a larger sample.

[6] https://utrl.ff.cuni.cz/

[7] Such choices are also likely to be changed in paraphrasing, which opens a new view on the results of Freitag et al. (2020), who show that using paraphrases as references in BLEU may lead to higher correlation with human evaluation.

Acknowledgments

The work was supported by the grants 19-26934X (NEUREM3) and GX20-16819X (LUSyD) by the Czech Science Foundation. The work has been using language resources developed and distributed by the LINDAT/CLARIAHCZ project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2018101).

References

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Ondřej Klejch, Eleftherios Avramidis, Aljoscha Burchardt, and Martin Popel. 2015. MT-ComparEval: Graphical evaluation interface for Machine Translation development. *The Prague Bulletin of Mathematical Linguistics*, 104:63–74.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Martin Popel. 2020. CUNI English-Czech and English-Polish Systems in WMT20: Robust Document-Level Training. In *Proceedings of the Fifth Conference on Machine Translation*, pages 269–273, Online. Association for Computational Linguistics.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(4381):1–15.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. ACL.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Gideon Toury. 1995. *Descriptive translation studies and beyond*. John Benjamins Publishing.