Machine Translation Human Evaluation: an investigation of evaluation based on Post-Editing and its relation with Direct Assessment

Luisa Bentivogli[(1)], *Mauro Cettolo*[(1)], *Marcello Federico*[(2)], *Christian Federmann*[(3)]

[(1)] FBK - Trento, Italy
[(2)] Amazon AI - East Palo Alto, CA, USA
[(3)] Microsoft Cloud+AI - Redmond, WA, USA

[bentivo|cettolo]@fbk.eu
marcfede@amazon.com
chrife@microsoft.com

Abstract

In this paper we present an analysis of the two most prominent methodologies used for the human evaluation of MT quality, namely evaluation based on Post-Editing (PE) and evaluation based on Direct Assessment (DA). To this purpose, we exploit a publicly available large dataset containing both types of evaluations. We first focus on PE and investigate how sensitive TER-based evaluation is to the type and number of references used. Then, we carry out a comparative analysis of PE and DA to investigate the extent to which the evaluation results obtained by methodologies addressing different human perspectives are similar. This comparison sheds light not only on PE but also on the so-called reference bias related to monolingual DA. Also, we analyze if and how the two methodologies can complement each other's weaknesses.

1. Introduction

The evaluation of machine translation (MT) is of crucial importance and has a long research history. Both human and automatic evaluation have been explored extensively within the MT community, in the effort to find more and more suitable, efficient and reliable methods and metrics. Automatic metrics play a central role in the progress of the field and the improvement of MT quality over time. However, they represent a proxy for human evaluation which – despite being costly and time-consuming – is to be considered primary.

Among the various human evaluation methods that have been devised and tested along the years, currently two approaches have become well-established standards in the field, namely evaluation based on *Post-Editing* (PE) and evaluation based on *Direct Assessment* (DA).

In the *PE-based evaluation*, the MT outputs are post-edited, *i.e.* manually corrected, according to the source sentence (*bilingual* PE) or to an existing reference translation (*monolingual* PE). The original MT outputs are then evaluated against their post-edited versions through TER-based automatic metrics [1]. Relying on the post-edit instead of an independently created reference translation ensures that only true errors in the MT output are counted, and not those differences due to linguistic variation, which are accounted for by post-editors. PE has become the standard evaluation metric for the yearly evaluation campaign of the International Workshop of Spoken Language Translation since 2013 (IWSLT-2013) and is described in detail in [2].

The *DA-based* evaluation [3] consists of collecting human assessments of translation quality for single MT systems. Assessors see a candidate translation and a corresponding translation hint (*e.g.* the source text, a reference translation, or multimodal content) and are asked to assign a quality score from 0 to 100. DA has become the standard evaluation metric for the yearly Conference on Machine Translation (WMT) in 2017 [4]. Following the findings of WMT17, the main focus for DA is on semantic transfer (which corresponds to *adequacy*) while syntactic transfer (or *fluency*) has turned out to be less relevant. Traditionally, in the DA task MT quality is assessed according to a reference translation, without access to the source text. This is called *reference-based DA* (*DA-ref*). A problematic issue with DA-ref is its inherent dependence on reference translations, which can lead to *reference bias*, both in the form of giving an implicit boost to candidate translations which are very similar (*e.g.*, in syntax or lexical choice) to the corresponding reference text, or by penalizing good translations because of translation errors affecting the reference itself. To address the reference bias, *source-based DA* (*DA-src*) can be used, where translation quality is assessed directly according to the source text. DA-src has been tested on a large scale for the first time in the IWSLT 2017 evaluation campaign [5].

DA and PE are different and complementary methodologies, not only from the point of view of their design but also concerning their practical usage. First, the two evaluation

[(2)] Work conducted while this author was at FBK.

methods address different human perspectives. Indeed, while DA focuses on the generic assessment of overall translation quality, PE-based evaluation reflects a real application scenario – the integration of MT in Computer-Assisted Translation (CAT) tools – and directly measures the utility of a given MT output to translators. Furthermore, while DA is based only on human annotators, in PE an automatic component (*i.e.* TER) is applied to quantify the errors of the MT output. Finally, in terms of data collection DA is less costly then PE and thus more viable when used within the research scenario; however PE has the double advantage of *(i)* producing a set of additional reference translations, and *(ii)* being particularly suitable for performing fine-grained analyses of the MT systems, since it produces a set of edits pointing to specific translation errors [6, 7, 8].

Given the importance of human evaluation for MT improvement and the specific features of these two most prominent frameworks, we present an empirical analysis of these different methodologies as a contribution to their better understanding.

The analysis is conducted on the publicly available Human Evaluation dataset created as part of the IWSLT 2017 evaluation campaign [5]. The dataset covers two language directions, namely Dutch-to-German and Romanian-to-Italian. For each direction, it includes DA-src, DA-ref, and PE human evaluation data for nine different state-of-the art neural MT systems on the same 603 segments. DA evaluation was performed by linguists, while professional translators carried out the bilingual PE task. Besides making our study possible, the size, variety and high quality of this three-way evaluation dataset ensure sound empirical analyses and generalizable outcomes.

The main investigations presented in the paper are:

- *New analyses on PE data.* The availability of multiple post-edits allows us to investigate how sensitive TER-based evaluation is to the type (external *versus* post-edit) and number of references used, both in terms of reliability and informativeness of the evaluation;

- *New comparative analysis of PE and DA.* In this empirical comparison we investigate the extent to which the evaluation results obtained by methodologies addressing different human perspectives are similar. This investigation gives us insight not only on PE but also on the relations between DA-src and DA-ref. Also, we analyze if and how PE and DA can complement each other's weaknesses.

2. Related Work

Human Evaluation has always received a lot of attention in the field of MT and many methodologies have been devised and tested in different scenarios. The same holds for the two methods addressed in this paper.

PE-based evaluation was the focus of various studies [1, 9, 6] and was commonly employed in large-scale evaluation campaigns, such as IWSLT [2, 10, 11, 12, 5] and the MT Quality Estimation Task at WMT-2015 [13].

Also research on DA has been very active since its introduction as method for human evaluation of MT [3, 14]. Large-scale evaluations were carried out through DA-ref [4] and, more recently, also through DA-src [5, 15].

As specifically regards the impact of different numbers and types of post-edits in PE-based evaluation, a study on multiple references was presented in [16], but it did not target PE-based evaluation.

Concerning the issue of reference bias in DA-ref evaluation, it was examined in detail in [17], [18], and [19]. To this aim, [17] compares directly DA-src and DA-ref but on a very small dataset, not comparable to the one used in our investigation.

As regards the comparative analysis of DA and PE, correlation results between DA-ref and HTER for 9 language directions are presented in [19]. However, the evaluation data differs in many respects, making results not comparable. First, the dataset used in this paper includes both DA-ref and DA-src. Furthermore, PE data is made of multiple bilingual post-edits created by professional translators native in the target language and working in their professional CAT environment. On the contrary, the post-edits used to calculate HTER in [19] were created through monolingual post-editing, probably based on the same reference used to collect DA-ref judgments.

3. Evaluation Data

To perform our investigations on DA and PE we relied on the Human Evaluation dataset created as part of the IWSLT 2017 evaluation campaign [5]. The resource is publicly available at the WIT[3] website [20], where all IWSLT data and tools are released by the organizers of the campaign. [1]

The dataset is based on TED talks[2] and includes 603 sentences (around 10,000 source words), corresponding to the first half of ten different TED talks. It covers two language pairs, namely Dutch-German ($NlDe$) and Romanian-Italian ($RoIt$) which – belonging to two distinct families (West-Germanic and Romance, respectively) – show rather different characteristics.

For each language direction, evaluation data were collected for nine different state-of-the-art neural MT systems: three standard *bilingual* systems (*i.e.* a different system is created for each language direction) and six *multilingual* systems (*i.e.* one single system for multiple language directions), out of which three in the *zero-shot* condition (*i.e.* tested on language pairs that are not present in the training data). Furthermore, systems differ also for their architecture, since some of them implement *Recurrent Neural Networks*, while others are based on the *Transformer* model [21].[3]

[1] https://wit3.fbk.eu/show.php?release=2017-02&page=subjeval&texthead=Evaluation%20Data

[2] www.ted.com

[3] All details about the MT systems can be found in [22, 23, 24].

The MT systems were evaluated on all the 603 dataset sentences according to PE, source-based DA, and reference-based DA. Details on human evaluation data are given in the following.

3.1. Post-Editing data

This evaluation was carried out through *bilingual* post-editing: the outputs of the nine MT systems on the 603 test sentences were assigned to nine professional translators to be manually corrected directly according to the source sentence.

To ensure the soundness of the evaluation and cope with translators' variability, an equal number of outputs from each MT system was assigned randomly to each translator, in such a way that each translator had to post-edit all the sentences in the test set but only once.

The resulting PE data used in this study consists of nine new reference translations for each sentence of the test set. Each one of these references represents the *targeted reference* of the system output from which it was derived, while the post-edits of the other systems are available for evaluation as additional references. All details about data preparation and post-editing can be found in [2, 5].

In addition to the PE data, an external - independently created - reference was also available, for a total of ten references for each of the 603 sentences in the dataset.

3.2. Direct Assessment data

Both DA-src and DA-ref data were collected for all the MT system outputs on all the 603 test sentences employing bilingual linguists. To ensure the reliability of the human assessments, part of the collected data was used for quality control. Based on artificially degraded translation output—which should be scored worse than the corresponding candidate translation—it is possible to identify users who randomly assign scores without paying attention to the presented data and, thus, work unreliably. Only annotations from reliable annotators were used to compute the final system evaluation. Furthermore, as annotators may have different annotation behaviour, the collected scores (at least two for each sentence) were standardized into *z scores*, which capture the number of standard deviations a score is different from (*i.e.* better or worse than) the respective annotator's mean score. Then, z scores were averaged at segment and system level to determine the overall MT system quality as observed by all annotators.

4. Analysis of PE-based evaluation

As described in Section 1, evaluation via post-editing is based on TER, which measures the amount of editing that a human would have to perform to change an automatic translation so that it exactly matches a given reference translation. Since TER is an automatic metric that works on exact word matching, it is unable to distinguish differences between MT output and reference due to normal linguistic variation from those due to real MT errors.

For this reason the reference translations used in TER-based evaluation (as in all automatic evaluations) play a central role in determining its reliability and informativeness.

It is widely accepted that the most suitable reference to evaluate an MT system is its corresponding post-edit (targeted reference), since it is derived from that specific system and thus should differ from the MT output only with respect to the parts of it that are incorrect. External references are at the other hand of the spectrum, since they are manually generated by translating the source text from scratch, independently from any MT system output. A particular case of reference is the post-edit of an actual system output which is not the one under evaluation. In this case the reference represents one of the many possible translation options and can indeed differ from the evaluated MT output due to linguistic variation. However, being created starting from an MT output, it is possible that its peculiar features make it more suitable to MT evaluation. This type of reference is particularly interesting since it can be easily gathered, being a natural by-product of professional translation in the CAT framework. Finally, the usage of multiple references has often been investigated as a way to address the issue of acceptable linguistic variation, under the assumption that the more references the highest the reliability of the evaluation.

In this section we exploited the PE data – *i.e.* one external reference and nine post-edits created from the nine evaluated MT systems – to carry out different analyses aimed at understanding if and how TER-based evaluation is sensitive to the type and number of references used.

Depending on the reference(s) used in the analysis, we relied on different variants of TER, namely: *(i) Human-targeted TER* (HTER), where TER is computed between the machine translation and its post-edited version (targeted reference); *(ii) Multiple reference TER* (mTER), where TER is computed against the closest reference – *i.e.* the one which minimizes the number of edits – among all the available ones.

We empirically analyzed the impact of references in the evaluation from two different angles: *(i)* for each evaluated MT system, we investigated the specific contribution of each of the nine available post-edits to the mTER score of the system; *(ii)* for each language pair, we calculated how overall MT system performance (*i.e.* TER score) varies depending on the type and number of references used.

Figure 1 shows an example of the distribution of the identity of systems which originated the post-edits that were chosen as closest reference translation in the computation of mTER. Four *NlDe* systems are presented in the figure, among which three were post-edited (BL.lab1, SD.lab2, ZS.lab3) and one was not (SD.lab4), and is shown for comparison purposes. The same behaviour of the *NlDe* systems presented in the figure was observed also for the other *NlDe* systems as well as for the *RoIt* direction.

As expected, the peak occurs in correspondence of the post-edit of the system under evaluation. Looking at the cor-

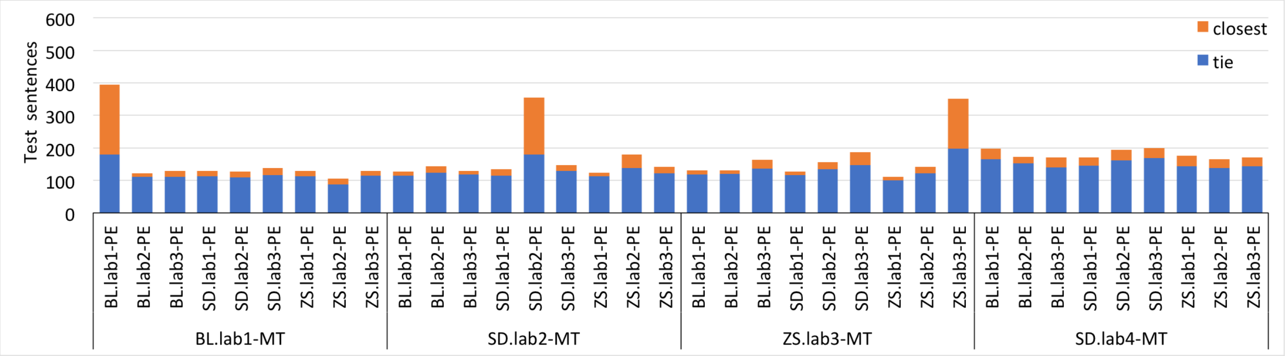

Figure 1: Frequency distribution of the closest PE selected in the computation of mTER of four *NlDe* systems. For each system that originated the PE, in orange the number of sentences for which that PE was the closest translation to the MT under investigation, in blue the number of sentences where the PE was the closest together with at least another PE.

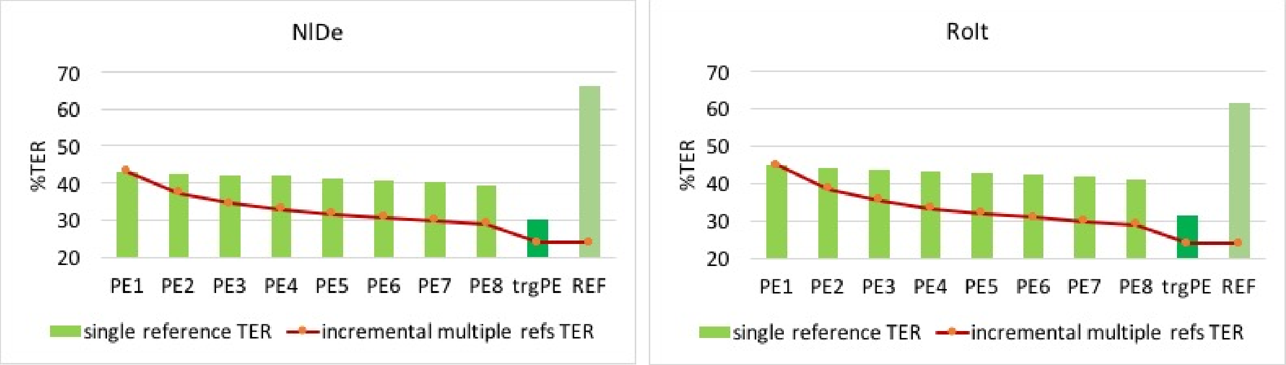

Figure 2: TERs on single references (green bars) and mTER on increasing number of references (red line).

responding column, we note however that the targeted reference is the closest to the MT output only for around one-third of the test set (orange-coloured), while for another third there is at least one equivalent post-edit from another system (blue-coloured). Interestingly enough, for the remaining third of the test set, the closest reference is a post-edit from another system.

Looking at the columns of the post-edits originated from the other 8 MT systems, we see that for a non-negligible number of test sentences these references represent the closest translation (orange). This is particularly relevant when confronted to the results of the external reference translation, which is not shown in the figure since it was never picked as the closest reference translation. It is also worthwhile to note that the post-edits of other MT systems created by the same Lab – which are expected to have similar outputs – are not chosen as closest references significantly more often than the post-edits of other Labs' systems. This suggests that the advantage of the post-edits of other systems does not rely in the similarity of the MT systems but more generally in the fact that the reference translation is derived from an MT output.

From the point of view of the number of references used in the evaluation, we understand from Figure 1 that a certain degree of variability is present also in the targeted translation – since for one-third of the test set it does not ensure the lowest edit distance with the MT output. We can thus confirm that – even when a targeted reference is available – mTER guarantees the highest reliability of the evaluation. Finally, the rightmost part of Figure 1 presents results for a system (SD.lab4) for which no post-edit was created. We can see that the closest references are equally distributed among all the available references, further confirming the importance of having multiple references.

The same conclusions can be drawn by analyzing the overall performances of the MT systems when using different reference translations. For each language direction, Figure 2 shows the impact that each of the ten references at our disposal has on TER, averaged across systems. The vertical bars provide the TER score computed using a single reference, be it one of the external post-edits, the targeted post-edit, or the external reference; for each system, the PEs are considered in reverse order with respect to their overall score, that is from the farthest to the closest to the system output, which invariably is the targeted PE; the external reference is presented as the last; the red line represents the mTER computed on an incremental set of references.

The low TER results obtained using a single non-targeted post-edit are quite interesting. Indeed evaluating a system against a post-edit created for another system is more sound than using an external reference. This is particularly relevant in a real application scenario where obtaining a post-edit of a system is easy and inexpensive. On the same line, considering the mTER cumulative score, it is interesting to see that the same HTER results obtained with the targeted reference (trgPE, dark green bar) can be achieved using seven external post-edits for the *NlDe* direction and six for the *RoIt* direction.

For completeness, Table 1 gives the exact figures of the most relevant information contained in Figure 2, namely mTER using all 9 available post-edits, HTER, and TER over the external reference.

Indeed we can observe a considerable TER reduction when using all collected post-edits with respect to both the HTER obtained using the targeted post-edit and the TER obtained using the independent reference. This reduction clearly confirms that exploiting all the available reference translations allows to produce a score which is not only more reliable but also more informative about the real performance of the systems.

	mTER 9 PE refs	**HTER** tgt PE	**TER** 1 ext ref
NlDe	23.80	29.96	66.10
RoIt	23.64	31.25	61.56

Table 1: %TERs computed on different (set of) references.

5. Comparative analysis of DA-based and PE-based evaluation

As introduced in Section 1, the DA-based and PE-based evaluation tasks focus on different aspects of automatic translation: general quality for the reader and usefulness for translator, respectively. To investigate the extent to which PE and DA lead to similar results, for each evaluated system we calculated the Pearson correlation between PE-based scores and DA-based scores for each sentence in the test set. The correlation results obtained for each system were then averaged through the Fisher transformations suggested in [25].

Table 2 presents the average correlation results. Correlations are calculated for both DA-src and DA-ref and for all the metrics investigated for PE-based evaluation, namely mTER, HTER and TER.

As expected, correlation is good, that is, in general segments judged as poor by DA annotators (low DA scores) also need substantial post-editing (high PE scores) or vice-versa.

Results slightly vary across language directions, but the same trends can be observed. First, the highest correlation is found between DA-src and mTER, confirming that these are the two most highly reliable human evaluation measures. As regards PE, mTER correlates better than HTER with DA, showing once again the importance of having multiple references. As regards DA, correlation with PE is considerably higher for DA-src than DA-ref. This indicates that the so-called reference-bias affects not only automatic metrics but also DA-based human evaluation. This is further confirmed by the results obtained for TER, which is calculated on the same external reference translation used in DA-ref. Although TER correlation scores are very low, TER correlates much better with DA-ref than DA-src, showing an opposite behaviour with respect to mTER and HTER.

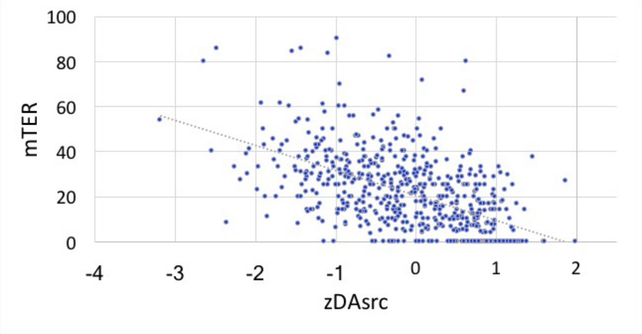

Figure 3: ZS.lab3 *RoIt* system: scatter plot of source-based DA standardized scores and mTER scores.

Given the correlation results obtained, we carried out a further analysis to investigate whether having both evaluations can help improving the evaluation quality, *i.e.* whether the two methodologies can complement each other's weaknesses.

Figure 3 shows the scatter plot of the correlation between mTER and DA-src for one of the investigated *RoIt* systems (r=–0.5812). Conflicting evaluations appear in the lower-left and upper-right quadrants of the scatter plot. The first quadrant includes segments which resulted good according to PE evaluation (low PE scores) but were judged as poor by DA annotators (low DA scores); the second includes segments which needed substantial post-editing (high PE scores) but were judged as good by DA annotators (high DA scores).

Conflicting evaluation cases are particularly relevant since PE is known to be more informative (see Section 1), but DA could identify issues that PE-based evaluation cannot spot. We manually inspected a sample of the sentences with conflicting evaluations and we found some interesting patterns. Examples are provided in Table 3.

When PE scores are low (*i.e.* few edits are needed to correct the MT output) but the translation is bad according to DA, typically the sentence contains few but crucial errors, which make it difficult to understand the meaning of the sentence (see Example 1 in the table). In these cases, the conflict is not solvable since from the point of view of DA - which is focused on adequacy - the MT output is rightfully not good, while from the point of view of the translator who has access to the source sentence, the MT output is indeed useful to speed-up translation.

In the opposite situation, *i.e.* high PE scores but good translation according to DA, we have two main causes for

avg(**r**)		*NlDe*			*RoIt*		
		mTER	HTER	TER	mTER	HTER	TER
zDA	src	-0.5466	-0.4796	-0.1918	-0.5294	-0.4306	-0.2137
	ref	-0.4491	-0.4100	-0.3579	-0.4524	-0.3882	-0.3570

Table 2: Average (DA,PE) correlations across systems.

			mTER	DA-src (abs)
1.	SRC	Nu are flapsuri, balamale, eleroane, actuatoare sau alte suprafee de control, doar o simplă elice. *It has no flaps, no hinges, no ailerons, no actuators, no other control surfaces, just a simple propeller.*		
	MT	Non ha **fiori**, **balconi**, **elenchi**, attuatori o altre superfici di controllo, solo una semplice elica. *It has no flowers, no balconies, no lists, no actuators, no other control surfaces, just a simple propeller.*	14.43%	28
	PE	Non ha **flaps**, **cerniere**, **alettoni**, attuatori o altre superfici di controllo, solo una semplice elica.		
2.	SRC	Prietenele mele, feministe convinse, au fost șocate. *My [female] friends, committed feminist, were aghast*		
	MT	**I miei** ami**ci**, femministe convint**i**, sono rimasti scioccat**i**. *My [male] friends, committed feminist, were aghast.*	47.87%	88
	PE	**Le mie** ami**che**, femministe convint**e**, sono rimasti scioccat**e**.		

Table 3: *RoIt* language direction. Examples of conflicting DA-PE evaluation.

conflicts. First, we found very short or long sentences which are indeed good translations but the mTER score was not correct due to tokenization (and consequently alignment) problems. These cases highlight the main weakness of PE-based evaluation, namely the fact that it relies on automatic metrics to compute the edit distance. The other type of conflict (see Example 2 in the table) regards those segments that have to be heavily post-edited for amending errors which do not alter the overall comprehension, like in chains of morphological errors. In these cases, the MT errors affect more fluency than adequacy, to which DA-based assessment is less sensitive.

6. Conclusions

In order to shed light on the properties, strengths and weaknesses of human evaluation it is crucial to rely on high quality datasets. The specific characteristics of the IWSLT-17 Human Evaluation dataset used in this investigation - size, variety and high quality of the three-way human evaluation - ensured sound empirical analyses and generalizable outcomes. The main findings of this paper are summarized in the following.

Analysis on PE evaluation data:

- the targeted reference is the closest to the MT output only for one-third of the test sentences. Thus, mTER guarantees the highest reliability of the evaluation over HTER;
- evaluating a system against a post-edit created for another system is more sound than using an external reference, independently from the similarity of the two MT systems;
- the same results obtained with the targeted reference (HTER) can be achieved using six/seven external post-edits (mTER), not including the targeted reference.

Comparative analysis of DA and PE:

- the highest correlation is found between DA-src and mTER, confirming that these are the two most highly reliable human evaluation measures;
- correlation with PE is considerably stronger for DA-src than DA-ref. This indicates that the so-called reference-bias affects not only automatic metrics but also DA-based human evaluation;
- conflicting evaluations between DA-src and mTER exist. In some cases DA-src can help mitigate the weakness of PE which depends on its automatic component. In other cases conflicts are caused by inherent differences due to the fact that the two evaluation methods address different human perspectives.

To conclude, we are planning to extend our research on both the analyses presented in this paper. First, we will further verify and generalize the results obtained on PE data by carrying out the analyses on other publicly available IWSLT datasets, which include multiple post-edits for other language directions such as English-German, English-French, and Vietnamese-English. Second, we will compare more deeply how DA-ref and DA-src behave on the same data. Finally, we will perform the manual analysis also on *NlDe* data.

7. References

[1] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proc. of AMTA*, Cambridge, US-MA, 2006, pp. 223–231.

[2] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and M. Federico, “Report on the 10th IWSLT evaluation campaign,” in *Proc. of IWSLT*, Heidelberg, Germany, 2013.

[3] Y. Graham, T. Baldwin, A. Moffat, and J. Zobel, “Continuous measurement scales in human evaluation of machine translation,” in *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Association for Computational Linguistics, 2013, pp. 33–41. [Online]. Available: http://www.aclweb.org/anthology/W13-2305

[4] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, S. Huang, M. Huck, P. Koehn, Q. Liu, V. Logacheva, C. Monz, M. Negri, M. Post, R. Rubino, L. Specia, and M. Turchi, “Findings of the 2017 conference on machine translation (wmt17),” in *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Copenhagen, Denmark: Association for Computational Linguistics, September 2017, pp. 169–214. [Online]. Available: http://www.aclweb.org/anthology/W17-4717

[5] M. Cettolo, M. Federico, L. Bentivogli, J. Niehues, S. Stüker, K. Sudoh, K. Yoshino, and C. Federmann, “Overview of the iwslt 2017 evaluation campaign,” in *Proc. of IWSLT*, Tokyo, Japan, 2017.

[6] M. Denkowski and A. Lavie, “Choosing the right evaluation for machine translation: An examination of annotator and automatic metric performance on human judgment tasks,” in *Proceedings of the 9th Conference of the Association of Machine Translation in the Americas (AMTA)*, Denver, CO, USA, 2010.

[7] L. Bentivogli, A. Bisazza, M. Cettolo, and M. Federico, “Neural versus phrase-based machine translation quality: a case study,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, November 2016, pp. 257–267. [Online]. Available: https://aclweb.org/anthology/D16-1025

[8] ——, “Neural versus phrase-based mt quality: An in-depth analysis on english-german and english-french,” *Computer Speech Language*, vol. 49, pp. 52 – 70, 2018.

[9] M. Snover, N. Madnani, B. J. Dorr, and R. Schwartz, “Fluency, adequacy, or hter?: exploring different human judgments with a tunable mt metric,” in *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2009, pp. 259–268.

[10] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and M. Federico, “Report on the 11th IWSLT evaluation campaign, IWSLT 2014,” in *Proc. of IWSLT*, Lake Tahoe, US-CA, 2014.

[11] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, R. Cattoni, and M. Federico, “The IWSLT 2015 evaluation campaign,” in *Proc. of IWSLT*, Da Nang, Vietnam, 2015.

[12] ——, “The IWSLT 2016 evaluation campaign,” in *Proc. of IWSLT*, Seattle, US-WA, 2016.

[13] O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, M. Huck, C. Hokamp, P. Koehn, V. Logacheva, C. Monz, M. Negri, M. Post, C. Scarton, L. Specia, and M. Turchi, “Findings of the 2015 workshop on statistical machine translation,” in *WMT@EMNLP*, 2015.

[14] Y. Graham, T. Baldwin, A. Moffat, and J. Zobel, “Is machine translation getting better over time?” in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014, pp. 443–451.

[15] H. Hassan, A. Aue, C. Chen, V. Chowdhary, J. Clark, C. Federmann, X. Huang, M. Junczys-Dowmunt, W. Lewis, M. Li, S. Liu, T. Liu, R. Luo, A. Menezes, T. Qin, F. Seide, X. Tan, F. Tian, L. Wu, S. Wu, Y. Xia, D. Zhang, Z. Zhang, and M. Zhou, “Achieving human parity on automatic chinese to english news translation,” *CoRR*, vol. abs/1803.05567, 2018. [Online]. Available: http://arxiv.org/abs/1803.05567

[16] A. Lommel, “Blues for bleu: Reconsidering the validity of reference-based mt evaluation,” in *Proceedings of the LREC 2016 Workshop Translation Evaluation-From Fragmented Tools and Data Sets to an Integrated Ecosystem*, 2016.

[17] M. Fomicheva and L. Specia, “Reference bias in monolingual machine translation evaluation,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2016, pp. 77–82. [Online]. Available: http://www.aclweb.org/anthology/P16-2013

[18] Q. Ma, Y. Graham, T. Baldwin, and Q. Liu, “Further investigation into reference bias in monolingual evaluation of machine translation,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017, pp. 2476–2485. [Online]. Available: http://aclweb.org/anthology/D17-1262

[19] Y. Graham, T. Baldwin, M. Dowling, M. Eskevich, T. Lynn, and L. Tounsi, “Is all that glitters in machine translation quality estimation really gold?”

in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, 2016, pp. 3124–3134. [Online]. Available: http://www.aclweb.org/anthology/C16-1294

[20] M. Cettolo, C. Girardi, and M. Federico, "WIT^3: Web inventory of transcribed and translated talks," in *Proc. of EAMT*, Trento, Italy, May 2012. [Online]. Available: http://hltshare.fbk.eu/EAMT2012/html/Papers/59.pdf

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[22] R. Dabre, F. Cromieres, and S. Kurohashi, "Kyoto university MT system description for IWSLT 2017," in *Proc. of IWSLT*, Tokyo, Japan, 2017.

[23] C. España-Bonet and J. van Genabith, "Going beyond zero-shot MT: combining phonological, morphological and semantic factors. The UdS-DFKI system at IWSLT 2017," in *Proc. of IWSLT*, Tokyo, Japan, 2017.

[24] S. M. Lakew, Q. F. Lotito, M. Turchi, M. Negri, and M. Federico, "FBKs multilingual neural machine translation system for IWSLT 2017," in *Proc. of IWSLT*, Tokyo, Japan, 2017.

[25] D. M. Corey, W. P. Dunlap, and M. J. Burke, "Averaging correlations: Expected values and bias in combined Pearson *r*s and Fisher's *z* transformations," *The Journal of General Psychology*, vol. 125(3):245-261, 1998.