Baltic J. Modern Computing, Vol. 4 (2016), No. 2, pp. 218–229

Potential and Limits of Using Post-edits as Reference Translations for MT Evaluation

Maja POPOVIĆ[1], Mihael ARČAN[2], Arle LOMMEL[3]

[1] Humboldt University of Berlin
[2] Insight Centre for Data Analytics, NUI Galway, Ireland
[3] Common Sense Advisory (CSA Research)

maja.popovic@hu-berlin.de, mihael.arcan@insight-centre.org, arle.lommel@gmail.com

Abstract. This work investigates the potential use of post-edited machine translation (MT) outputs as reference translations for automatic machine translation evaluation, focusing mainly on the following important question: *Is it necessary to take into account the machine translation system and the source language from which the given post-edits are generated?*
In order to explore this, we investigated the use of post-edits originating from different machine translation systems (two statistical systems and two rule-based systems), as well as the use of post-edits originating from two different source languages (English and German). The obtained results shown that for comparison of different systems using automatic evaluation metrics, a good option is to use a post-edit originating from a high-quality (possibly distinct) system. A better option is to use it together with other references and post-edits, however post-edits originating from poor translation systems should be avoided. For tuning or development of a particular system, post-edited output of this same system seems to be the best reference translation.

Keywords: machine translation evaluation, reference translations, post-edited translations

1 Introduction

The evaluation of the machine translation (MT) output is an important and difficult task. The fastest way is to use an automatic evaluation metric, which compares the obtained output with a human translation of the same source text and calculates a numerical score related to their similarity. Despite all disadvantages and criticisms, such metrics are still irreplaceable for many tasks (such as rapid development of a new system, tuning of a statistical MT system, etc.) and are considered as at least baseline metrics for MT quality evaluation. All these metrics (n-gram based such as BLEU [Papineni et al., 2002] and METEOR [Banerjee and Lavie, 2005], edit-distance based such as TER [Snover et al., 2006], etc.) are reference-based, i.e. a human reference translation is needed as a gold standard. Since there is usually not only one single best translation of a text, the best way of evaluating an MT output would be to compare it with many references

– nevertheless, creating each reference translation is a time consuming and expensive process. Therefore, automatic MT evaluation is usually carried out using only a single reference.

On the other hand, MT has considerably improved in the recent years so that the use of MT outputs as a starting point for human translation has become a common practice. Therefore, ever-increasing amounts of post-edited machine translation outputs (PEs) are being collected. These represent very valuable data and are being used for a number of applications, such as automatic quality prediction, adaptation, etc. Among other things, post-edits are more similar to MT outputs than "independent" references, thus being potentially more useful for automatic evaluation and/or tuning. However, their use as reference translations has been scarcely investigated so far.

This work explores two scenarios: comparing four distinct MT systems using PEs originating from these systems, as well as comparing translations from two different source languages using PEs originating from these source languages. In addition, the effects of using multiple references are reported in terms of variations and standard deviations of automatic scores for different number of references.

1.1 Related work

Post-edited translations have been used for many applications, such as automatic prediction of translation quality [Specia, 2011], analysing various aspects of post-editing effort [Tatsumi and Roturier, 2010, Blain et al., 2011], human and automatic analysis of performed edit operations [Koponen, 2012, Wisniewski et al., 2013], as well as improving translation and language model of an SMT system by learning from post-edits [Bertoldi et al., 2013, Denkowski et al., 2014, Mathur et al., 2014]. The cache-based approach, introduced in [Bertoldi et al., 2013], makes it possible to periodically add knowledge from PEs into an SMT system in real-time, without the need to stop it. The main idea behind the cache-based models is to mix a large global (static) model with a small local (dynamic) model estimated from recent items observed in the history of the input stream. In [Wisniewski et al., 2013], the PEs are used as references for automatic estimation of performed edit operations, namely substitutions, deletions, insertions and shifts. [Denkowski et al., 2014] report the improvements of the BLEU scores calculated on independent references as well as on PEs in order to emphasise the suitability of their methods for the post-editing task.

A number of publications deals with the usage of multiple references for automatic MT evaluation. Using pseudo-references, i.e. raw translation outputs from different MT systems has been investigated in [Albrecht and Hwa, 2007, Albrecht and Hwa, 2008] and it is shown that, even though these are not correct human translations, it is beneficiary to add pseudo-references instead of using one single reference. Adding automatically generated paraphrases together to a set of standard human references for tuning has been investigated in [Madnani et al., 2008], and it is shown that the paraphrases are improving automatic scores BLEU and TER when the number of multiple human references is less than four. Recently, multiple references have been explored in [Qin and Specia, 2015] in terms of using recurring information in these references in order to generate better version of BLEU and NIST [Doddington, 2002] metrics by better n-gram weighting.

To the best of our knowledge, no systematic investigation regarding the use of post-edited translation outputs as reference translations has been carried out yet.

2 Research questions

Although the PEs are intuitively better suitable for MT evaluation than standard human references because they are closer to the MT output structure, there are several important questions which have to be taken into account:

1. Should the PE originate from the very same MT system, or is it acceptable to use any PE?
2. Is the source language of any importance?
3. Does the system type (statistical or rule-based) have any impact?

In order to systematically explore the potential and limits of post-edits and answer these questions, following scenarios are investigated:

- using PEs produced by four distinct MT systems;
- using PEs generated from two different source languages;

The PEs are used for system comparison in order to explore variations and possible bias of the obtained automatic scores. Apart from the use of each post-edit separately, the effects of combining them in the form of multiple references has been investigated. In addition, the effect of the source language has also been explored in terms of tuning an SMT system. The details about the experiments and the obtained results are described in the next two sections.

3 Experiments

3.1 Data sets

For investigation of effects described in the previous section, two suitable data sets containing different language pairs, target languages and domains were available:

1. TARAXÜ texts [Avramidis et al., 2014] containing German-to-English, German-to-Spanish and English-to-German raw translations and PEs of WMT news texts generated by two SMT (phrase-based and hierarchical) and two RBMT systems;
2. OPENSUBTITLES texts from the PE^{2}rr corpus [Popović and Arčan, 2016] containing Serbian and Slovenian subtitle raw translations and PEs generated by phrase-based SMT systems from English and from German.

Both data sets contain single standard reference translations, as well as sentence-level human rankings.

For the TARAXÜ WMT texts, post-editing and ranking were performed by professional translators, and for the PE^{2}rr OPENSUBTITLES texts by researchers familiar with machine and human translation highly fluent both in source and in target languages. Details about the texts can be seen in Table 1.[4]

[4] Although the texts are already publicly available, they are also available in the exact form used in this work at `https://github.com/m-popovic/multiple-edits-refs`.

Table 1: Data statistics

domain	language pair	# source sentences	avg. target sent. length	# of PE
WMT	de-en	240	22.9	4
(TARAXÜ)	de-es	40	26.8	4
	en-de	272	21.9	4
	es-de	101	23.2	4
OPEN	en-sr	440	8.3	2
SUBTITLES	de-sr	440	8.1	2
(PE^2rr)	en-sl	440	8.7	2
	de-sl	440	8.5	2

It should be noted that, although there are more (larger) publicly available data sets containing post-edited MT outputs, none of these sets contains post-edits originating from different translation systems or from different source languages, which are requested to answer the questions posed in Section 2.

3.2 Evaluation methods

For all experiments, BLEU scores [Papineni et al., 2002] and character n-gram F scores, i.e. CHRF3 scores [Popović, 2015], calculated using different PEs are reported. BLEU is used as a well-known and widely used metric, and CHRF3 as a simple tokenisation-independent metric, which has shown very good correlations with human judgements on the WMT-2015 shared metric task [Stanojević et al., 2015], both on the system level as well as on the segment level, especially for morphologically rich(er) languages.

For both scores, Pearson's system-level correlation coefficient r is reported for each PE. For CHRF3, segment-level Kendall's τ correlation coefficient is presented as well. For both correlation coefficients, the ties in human rankings were excluded from calculation. In all tables, post-edited MT outputs are marked with ^{pe}.

Initially, for each of the two data sets the scores were calculated separately for each target language. Nevertheless, since no differences related to the target language were observed, the results were merged.

4 Results

4.1 Post-edits from (four) different translation systems

In order to investigate PEs originating from different MT systems, the TARAXÜ corpus was used, where each source sentence was translated by four MT systems. Although there is certain overlap, i.e. some of the source sentences are human translations of other source sentences, the majority of them are unique. BLEU and CHRF3 scores are calculated separately using each of PEs as reference, as well as for combinations of

multiple PEs. The scores, together with system-level and segment-level correlation coefficients, are presented in Table 2.

Table 2: BLEU (left) and CHRF3 (right) scores calculated on PEs originating from four distinct MT systems (two SMT and two RBMT) and on an independent reference translation; the scores are strongly biased towards the particular system and slightly biased towards the system type; the best option is to use PE of a high performance system or multiple references without PEs of poor quality systems.

BLEU scores	translation output				corr.
# reference(s)	S1	S2	RB1	RB2	sys
1 S1^{pe}	**41.0**	25.8	22.5	20.0	-.40
S2^{pe}	27.8	**35.4**	21.2	19.7	-.99
RB1^{pe}	22.4	19.4	**46.6**	25.9	.75
RB2^{pe}	21.7	19.4	28.8	**41.3**	.77
reference	12.7	11.4	12.0	10.6	-.15
2 two SMT^{pe}	43.4	38.0	27.0	24.7	-.97
two RBMT^{pe}	27.0	23.6	48.7	43.5	**.99**
3 no S1^{pe}	34.8	38.1	49.5	44.4	.93
no S2^{pe}	43.8	31.1	49.5	44.5	**.98**
no RB1^{pe}	44.3	38.9	35.6	43.9	.20
no RB2^{pe}	44.4	38.8	48.9	32.2	.19
4 all^{pe}	44.9	39.4	50.0	45.1	.92
5 all^{pe}+ref	46.5	40.8	51.6	45.8	.88
human ranks	57.6	47.6	69.3	67.4	

CHRF3 scores	translation output				corr.	
# reference(s)	S1	S2	RB1	RB2	sys	seg
1 S1^{pe}	**67.9**	55.8	55.7	54.4	-.24	.03
S2^{pe}	58.1	**63.5**	54.6	53.8	-.98	.12
RB1^{pe}	54.0	50.7	**72.3**	58.6	.83	.29
RB2^{pe}	53.3	50.6	59.4	**69.5**	.80	.24
reference	43.4	41.4	44.0	43.6	.93	.13
2 two SMT^{pe}	68.6	64.2	58.1	57.0	-.71	.30
two RBMT^{pe}	56.6	53.4	72.8	70.2	**.96**	**.34**
3 no S1^{pe}	60.9	64.3	73.0	70.4	.78	**.38**
no S2^{pe}	68.6	58.0	73.0	70.4	**.95**	.30
no RB1^{pe}	68.8	64.4	62.5	70.2	.10	.14
no RB2^{pe}	68.9	64.5	72.9	61.4	.26	.20
4 all^{pe}	69.0	64.7	73.2	70.6	**.97**	.30
5 all^{pe}+ref	69.1	64.8	73.2	70.6	**.97**	.30
human ranks	32.0	22.0	54.8	46.5		

The following can be observed:

- each system gets the highest score when its own PE is used as a reference (bold); system level correlations are very low if the worse ranked system's PEs are used – in such scenario, worst systems obtain the highest automatic scores;
- the scores for both of SMT systems are higher if the two SMTPEs are used; analogously applies for the RBMT systems;
- the best options in terms of correlations are
  - using PE of the best ranked system;
  - not using PE of the worst ranked system;
  - using all PEs (and reference).

Table 3 presents edit distances between PEs as well as between PEs and the reference, and it can be seen that the differences are not negligible, which explains the strong bias towards the particular system. It can also be seen that the post-edits of the same system types are slightly closer (∼35%) than those of the two different system types

Table 3: Edit distances between PEs originating from four distinct systems and reference; the PEs of the same system types are slightly closer than those of the two different system types; the reference is significantly different from all PEs.

edit distance	S1^{pe}	S2^{pe}	RB1^{pe}	RB2^{pe}	ref
S1^{pe}	/	34.3	41.0	42.9	70.0
S2^{pe}	34.2	/	41.9	42.4	70.1
RB1^{pe}	40.5	41.5	/	35.4	70.9
RB2^{pe}	41.7	41.3	34.7	/	69.6
ref	69.0	69.2	70.6	70.5	/

(∼42%), as well as that there is a large distance (∼70%) between the reference and each of the PEs.

An example of German-to-English translation outputs, PEs and the corresponding reference is presented in Table 4.

Table 4: Example of post-edited German-to-English MT outputs originating from four distinct translation systems.

system	translation output	PE
S1	There are also a few cars off the road.	There are also a few cars off the road.
S2	Few cars are off the road.	A few cars are also off the road.
RB1	Also a few Pkws lie in the street ditch.	Also, a few cars are lying on the side of the street.
RB2	Also a few car lies in the ditch.	A few cars are also lying in the ditch.
	reference:	Also several cars ended up in a ditch.

4.2 Post-edits from (two) different source languages

For exploring influence of the source language, the OPENSUBTITLES texts were used, where each of the parallel German and English source sentences was translated by a corresponding phrase-based SMT system. The effects of the source language on the automatic scores are shown in Table 5. Since there are only two systems to compare, system-level Pearson's correlation coefficient can be either 1 or -1.

It can be noted that:

– the source language strongly influences the results: for each translation output, the automatic scores are always higher when its own PE is used;

Table 5: BLEU (left) and CHRF3 (right) scores calculated on SMTPEs originating from two different source languages and on an independent reference translation; the results are strongly biased towards the source language; the best option is to use PE of a high performance system or multiple references without PEs of poor systems.

	BLEU scores	translation output		corr.
	reference(s)	en→x	de→x	sys
1	en→x^{pe}	**47.7**	23.9	1
	de→x^{pe}	24.4	**45.5**	-1
	reference	24.8	17.2	1
2	en→x^{pe}+ref	51.3	28.2	1
	de→x^{pe}+ref	35.9	47.9	-1
	both^{pe}	50.4	48.0	1
3	both^{pe}+ref	53.0	49.2	1
	human ranks	38.6	21.1	

	CHRF3 scores	translation output		corr.	
	reference(s)	en→x	de→x	sys	seg
1	en→x^{pe}	**64.7**	44.6	1	.42
	de→x^{pe}	45.8	**62.6**	-1	.13
	reference	47.8	39.4	1	.44
2	en→x^{pe}+ref	65.6	47.8	1	.42
	de→x^{pe}+ref	54.6	63.4	-1	.28
	both^{pe}	66.5	63.8	1	.48
3	both^{pe}+ref	67.2	64.3	1	**.50**
	human ranks	38.6	21.1		

- using PE of the better ranked system yields good correlation, whereas using PE from the worse system claims that this system is better;
- the scores obtained by the independent reference are more similar to those obtained by the PE generated from English.

Furthermore, Table 6 shows that edit distances between PEs are rather large, about 45%. Similar edit distance can be seen between the independent reference and the PE originating from English, whereas for the PE originating from German it is much larger – over 55%. At this point, it is important to note that the original source language of all used texts is English – the German source text as well as the Serbian and Slovenian references are human translations of the English original. Therefore, the fact that the PE originating from German source is an "outlier" confirms the previous findings about the importance of the original source language, e.g. [Kurokawa et al., 2009, Lembersky et al., 2013], namely that (i) a translated text has different characteristics than the same text written directly in the given language, as well as that (ii) the direction of human translation has impact on MT, so that it is better to train MT system in the corresponding direction, i.e. using original texts as the source language and human translations as the target language.

4.3 Multiple reference effects

Apart from the main questions posed in Section 2, an additional question has been raised during the realisation of the described experiments – what are the actual effects of the use of multiple references vs. the use of a single reference?

The advantage of multiple references is surely well known as mentioned in Section 1.1, however our question is – what is exactly happening with the automatic scores? In order to answer it, we explored the variations in automatic scores when different

Table 6: Edit distances between post-edits originating from two different source languages and an independent reference translation.

edit distance	en-x^{pe}	de-x^{pe}	reference
en→x^{pe}	0	44.6	44.7
de→x^{pe}	45.7	0	56.4
reference	45.2	55.6	0

numbers of multiple references are used. For this experiment, apart from the two data sets described in previous sections, an additional small data set[5] was explored as well. This data set consists of only 20 English source sentences from technical domain, however each source sentence corresponds to 12 different human translations into German, i.e. 12 multiple references. Each source sentence has been automatically translated by four distinct translation systems, two statistical and two rule-based (albeit not the same as those used for experiments in Section 4.1), but no post-editing has been performed.

For each of the three data sets, average BLEU and CHRF3 values and their standard deviations (σ) for different numbers of available reference translations are calculated and results are presented in Table 7. It can be seen that:

- average values are logarithmically increasing with increasing number of multiple references;
- standard deviations are
  - dropping with increasing number of multiple references
  - close to zero only for more than 10 references
  - smaller for the MT systems of lower performance

These tendencies can be equally observed for all data sets, no matter how many PEs (more similar to MT outputs) and how many independent references (less similar to MT outputs) are used.

4.4 Tuning

A preliminary experiment regarding tuning on PEs originating from different source languages has been carried out using the OPENSUBTITLES data set: (i) the translation system was tuned with MERT [Och, 2003] on BLEU using (i) the independent reference (standard method), (ii) using the PE originating from the corresponding language and (iii) using the PE originating from the other language.

The results for another test set (not the one used for tuning) containing 2000 sentences[6] are presented in Table 8 showing that tuning on the post-edit from the corresponding source language produces best BLEU and METEOR scores.

This confirms the effect of the source language bias and indicates a potential of using PEs of a MT system for tuning and development of this system.

[5] also available at `https://github.com/m-popovic/multiple-edits-refs`
[6] also available at the aforementioned repository

Table 7: Effects of the number of multiple references: average BLEU and CHRF3 scores with standard deviations for different number of (independent) references ranging from 1 to 12. The results are obtained on the texts used in previous sections (a), (b) as well as on a small text with a large number (12) of independent reference translations (c).

(a) PEs of four different systems + one reference

		translation output							
	number of references	SMT1 avg.	SMT1 σ	SMT2 avg.	SMT2 σ	RBMT1 avg.	RBMT1 σ	RBMT2 avg.	RBMT2 σ
BLEU	1	25.1	9.3	22.3	8.0	26.2	11.5	23.5	10.2
	2	35.0	7.0	30.8	5.9	37.3	9.5	33.3	8.3
	3	40.3	5.3	35.4	4.4	43.6	7.6	38.8	6.7
	4	43.9	3.4	38.5	2.8	48.2	5.2	42.8	4.6
	all (5)	46.5	/	40.8	/	51.6	/	45.8	/
CHRF3	1	55.3	7.9	52.4	7.2	57.2	9.1	56.0	8.4
	2	62.0	5.7	58.4	5.0	64.6	7.0	62.9	6.2
	3	65.1	4.5	61.3	3.8	68.3	5.7	66.2	5.0
	4	67.4	2.9	63.3	2.6	71.0	4.2	68.7	3.5
	all (5)	69.1	/	64.8	/	73.2	/	70.6	/

(b) PEs of two source languages + one reference

		translation output			
	number of references	en→x avg.	en→x σ	de→x avg.	de→x σ
BLEU	1	32.3	10.9	28.9	12.1
	2	45.9	7.0	41.4	9.3
	all (3)	53.0	/	49.2	/
CHRF3	1	52.8	8.5	48.9	9.9
	2	62.2	5.4	58.3	7.4
	all (3)	67.2	/	64.3	/

(c) twelve references

		translation output							
	number of references	sys1 avg.	sys1 σ	sys2 avg.	sys2 σ	sys3 avg.	sys3 σ	sys4 avg.	sys4 σ
BLEU	1	32.1	8.4	29.4	9.7	23.2	6.5	13.3	5.0
	2	41.6	6.0	39.2	9.2	29.2	4.2	17.7	2.9
	10	61.2	1.7	62.4	2.0	40.5	0.6	26.0	0.6
	11	62.0	1.2	63.3	1.3	40.8	0.4	26.3	0.4
	all (12)	62.8	/	64.0	/	41.1	/	26.6	/
CHRF3	1	63.6	7.8	61.5	8.6	56.0	5.6	54.2	5.7
	2	71.1	5.1	69.3	6.8	62.0	3.3	60.0	3.7
	10	80.5	0.5	81.3	1.1	68.6	0.3	67.1	0.3
	11	80.7	0.3	81.7	0.6	68.8	0.2	67.3	0.2
	all (12)	81.0	/	82.0	/	69.0	/	67.5	/

Table 8: Effects of source language on tuning of an SMT system: MERT tuning on BLEU using independent reference, post-edit from the corresponding source language and post-edit from another source language. The best BLEU and METEOR scores are obtained when the corresponding source language post-edit is used.

translating	tuned on	BLEU	METEOR	translating	tuned on	BLEU	METEOR
en→sr	ref	20.1	39.2	en→sl	ref	26.0	45.2
	en→sr^{pe}	**21.6**	**39.9**		**en→sl^{pe}**	**26.5**	**45.3**
	de→sr^{pe}	20.8	39.8		de→sl^{pe}	25.5	44.4
de→sr	ref	17.2	35.4	de→sl	ref	18.1	36.6
	en→sr^{pe}	16.8	35.5		en→sl^{pe}	18.4	36.7
	de→sr^{pe}	**18.0**	**35.5**		**de→sl^{pe}**	**18.8**	**37.1**

5 Discussion

Knowing how difficult the generation of (even a single) references/PEs is, the following findings from the results described in Section 4 can be summarised:

- for comparison of different systems, using single PE of a high quality translation output yields reliable automatic scores; the scores are even more reliable if the PE is generated by an external system – otherwise, the ranking would be still correct but the scores will be biased to this particular system;
- using multiple PEs (and references) is generally beneficial – however, it is better to have fewer PEs of high quality translation outputs than more PEs of low quality translation outputs;
- evaluation of low quality translation outputs is less prone to variability and is generally more reliable, except if (one of) the used reference(s) is its own PE; on the other hand, high quality translation outputs can easily be underestimated if using a single reference/PE;
- for tuning and development of a particular system, the PE from this very system should be used.

6 Summary and outlook

This work has examined the potential and limits of the use of post-edited MT outputs as reference translations for automatic MT evaluation. The experiments have shown that the post-edited translation outputs are definitely useful as reference translations, but it should be kept in mind that the obtained automatic evaluation scores are strongly biased towards the actual system by which the used PE is generated, as well as towards the source language from which the used PE originates. The best option for comparison of different systems using a single PE is to use PE of a high quality translation output which is, if possible, generated by an independent system.

Multiple references are in principle beneficial, although PEs generated from low quality translation outputs should be avoided. Further investigation concerning both quality and quantity of multiple references should be carried out.

For tuning an SMT system, the best option is to use a PE generated by this same system. Nevertheless, it should be noted that this was a preliminary experiment, so that further confirmation of reported findings on more data and language pairs is necessary.

Acknowledgments

This publication has emanated from research supported by the TRAMOOC project (Translation for Massive Open Online Courses), partially funded by the European Commission under H2020-ICT-2014/H2020-ICT-2014-1 under Grant Agreement Number 644333, and by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (Insight)

References

Joshua S. Albrecht, Rebecca Hwa (2007). Regression for Sentence-level MT Evaluation with Pseudo-references. In *P*roceedings of the 45rd Annual Meeting of the Association for Computational Linguistics (ACL-07), pages 296–303, Prague, Czech Republic, July.

Joshua S. Albrecht and Rebecca Hwa. 2008. The Role of Pseudo-references in MT evaluation. In *Proceedings of the 3rd Workshop on Statistical Machine Translation (WMT-08)*, pages 187–190, Columbus, Ohio, June.

Eleftherios Avramidis, Aljoscha Burchardt, Sabine Hunsicker, Maja Popović, Cindy Tscherwinka, David Vilar Torres, and Hans Uszkoreit. 2014. The taraXÜ Corpus of Human-Annotated Machine Translations. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-14)*, pages 2679–2682, Reykjavik, Iceland, May.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgements. In *Proceedings of the ACL-05 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72, Ann Arbor, MI, June.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. Cache-based Online Adaptation for Machine Translation Enhanced Computer Assisted Translation. In *Proceedings of MT Summit XIV*, Nice, France.

Frédéric Blain, Jean Senellart, Holger Schwenk, Mirko Plitt, and Johann Roturier. 2011. Qualitative Analysis of Post-Editing for High Quality Machine Translation. In *Proceedings of Machine Translation Summit XIII*, Xiamen, China, September.

Michael Denkowski, Chris Dyer, and Alon Lavie. 2014. Learning from Post-Editing: Online Model Adaptation for Statistical Machine Translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL-14)*, pages 395–404, Gothenburg, Sweden, April.

George Doddington. 2002. Automatic Evaluation of Machine Tanslation Quality using n-gram Co-occurrence Statistics. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 128–132, San Diego, CA, March.

Maarit Koponen. 2012. Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the 7th Workshop on Statistical Machine Translation (WMT-12)*, pages 181–190, Montréal, Canada, June.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic Detection of Translated Text and its Impact on Machine Translation. In *Proceedings of MT Summit XII*, pages 81–88, Ottawa, Canada, August.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2013. Improving Statistical Machine Translation by Adapting Translation Models to Translationese. *Computational Linguistics*, 39(4):999–1024, December.

Nitin Madnani, Philip Resnik, Bonnie J. Dorr, and Richard Schwartz. 2008. Are multiple reference translations necessary? Investigating the value of paraphrased reference translations in parameter optimization. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas (AMTA-08)*, Waikiki, Hawaii, October.

Prashant Mathur, Mauro Cettolo, Marcello Federico, and José Guillerme Carmago de Souza. 2014. Online Multi-User Adaptive Statistical Machine Translation. In *Proceedings of the 11th Conference of the Association for Machine Translation of the Americas (AMTA-14)*, pages 152–165, Vancouver, Canada, October.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics (ACL 03)*, pages 160–167, Sapporo, Japan, July.

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 311–318, Philadelphia, PA, July.

Maja Popović. 2015. chrF: Character n-gram F-score for Automatic MT Evaluation. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT-15)*, pages 392–395, Lisbon, Portugal, September.

Maja Popović, Mihael Arčan. 2016. PE2rr corpus: Manual Error Annotation of Automatically Pre-annotated MT Post-edits. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-16)*, Portorož, Slovenia, May.

Ying Qin and Lucia Specia. 2015. Truly Exploring Multiple References for Machine Translation Evaluation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT-15)*, pages 113–120, Antalya, Turkey, May.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Error Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-06)*, pages 223–231, Boston, MA, August.

Lucia Specia. 2011. Exploiting Objective Annotations for Measuring Translation Post-editing Effort. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT-11)*, pages 73–80, Leuven, Belgium, May.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 Metrics Shared Task. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT-15)*, pages 256–273, Lisbon, Portugal, September.

Midori Tatsumi and Johann Roturier. 2010. Source Text Characteristics and Technical and Temporal Post-editing Effort: What is their Relationship? In *Proceedings of the Second Joint EM+/CGNL Worskhop Bringing MT to the user (JEC-10)*, pages 43–51, Denver, Colorado, November.

Guillaume Wisniewski, Anil Kumar Singh, Natalia Segal, and François Yvon. 2013. Design and Analysis of a Large Corpus of Post-edited Translations: Quality Estimation, Failure Analysis and the Variability of Post-edition. In *Proceedings of MT Summit XIV*, pages 117–124, Nice, France, September.

Received May 2, 2016 , accepted May11, 2016