Reference Bias in Monolingual Machine Translation Evaluation

Marina Fomicheva
Institute for Applied Linguistics
Pompeu Fabra University, Spain
marina.fomicheva@upf.edu

Lucia Specia
Department of Computer Science
University of Sheffield, UK
l.specia@sheffield.ac.uk

Abstract

In the translation industry, human translations are assessed by comparison with the source texts. In the Machine Translation (MT) research community, however, it is a common practice to perform quality assessment using a reference translation instead of the source text. In this paper we show that this practice has a serious issue – annotators are strongly biased by the reference translation provided, and this can have a negative impact on the assessment of MT quality.

1 Introduction

Equivalence to the source text is the defining characteristic of translation. One of the fundamental aspects of translation quality is, therefore, its semantic adequacy, which reflects to what extent the meaning of the original text is preserved in the translation. In the field of Machine Translation (MT), on the other hand, it has recently become common practice to perform quality assessment using a human reference translation instead of the source text. Reference-based evaluation is an attractive practical solution since it does not require bilingual speakers.

However, we believe this approach has a strong conceptual flaw: the assumption that the task of translation has a single correct solution. In reality, except for very short sentences or very specific technical domains, the same source sentence may be correctly translated in many different ways. Depending on a broad textual and real-world context, the translation can differ from the source text at any linguistic level – lexical, syntactic, semantic or even discourse – and still be considered perfectly correct. Therefore, using a single translation as a proxy for the original text may be unreliable.

In the monolingual, reference-based evaluation scenario, human judges are expected to recognize acceptable variations between translation options and assign a high score to a good MT, even if it happens to be different from a particular human reference provided. In this paper we argue that, contrary to this expectation, annotators are strongly biased by the reference. They inadvertently favor machine translations (MTs) that make similar choices to the ones present in the reference translation. To test this hypothesis, we perform an experiment where the same set of MT outputs is manually assessed using different reference translations and analyze the discrepancies between the resulting quality scores.

The results confirm that annotators are indeed heavily influenced by the particular human translation that was used for evaluation. We discuss the implications of this finding on the reliability of current practices in manual quality assessment. Our general recommendation is that, in order to avoid reference bias, the assessment should be performed by comparing the MT output to the original text, rather than to a reference.

The rest of this paper is organized as follows. In Section 2 we present related work. In Section 3 we describe our experimental settings. In Section 4 we focus on the effect of reference bias on MT evaluation. In Section 5 we examine the impact of the fatigue factor on the results of our experiments.

2 Related Work

It has become widely acceptable in the MT community to use human translation instead of (or along with) the source segment for MT evaluation. In most major evaluation campaigns (ARPA (White et al., 1994), 2008 NIST Metrics for Machine Translation Challenge (Przybocki et al., 2008), and annual Workshops on Statistical Ma-

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 77–82,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

chine Translation (Callison-Burch et al., 2007; Bojar et al., 2015)), manual assessment is expected to consider both MT fluency and adequacy, with a human (reference) translation commonly used as a proxy for the source text to allow for adequacy judgement by monolingual judges.

The reference bias problem has been extensively discussed in the context of automatic MT evaluation. Evaluation systems based on string-level comparison, such as the well known BLEU metric (Papineni et al., 2002) heavily penalize potentially acceptable variations between MT and human reference. A variety of methods have been proposed to address this issue, from using multiple references (Dreyer and Marcu, 2012) to reference-free evaluation (Specia et al., 2010).

Research in manual evaluation has focused on overcoming annotator bias, i.e. the preferences and expectations of individual annotators with respect to translation quality that lead to low levels of inter-annotator agreement (Cohn and Specia, 2013; Denkowski and Lavie, 2010; Graham et al., 2013; Guzmán et al., 2015). The problem of reference bias, however, has not been examined in previous work. By contrast to automatic MT evaluation, monolingual quality assessment is considered unproblematic, since human annotators are supposed to recognize meaning-preserving variations between the MT output and a given human reference. However, as will be shown in what follows, manual evaluation is also strongly affected by biases due to specific reference translations.

3 Settings

To show that monolingual quality assessment depends on the human translation used as gold-standard, we devised an evaluation task where annotators were asked to assess the same set of MT outputs using different references. As control groups, we have annotators assessing MT using the same reference, and using the source segments.

3.1 Dataset

MT data with multiple references is rare. We used MTC-P4 Chinese-English dataset, produced by Linguistic Data Consortium (LDC2006T04). The dataset contains 919 source sentences from news domain, 4 reference translations and MT outputs generated by 10 translation systems. Human translations were produced by four teams of professional translators and included editor's proofreading. All teams used the same translation guidelines, which emphasize faithfulness to the source sentence as one of the main requirements.

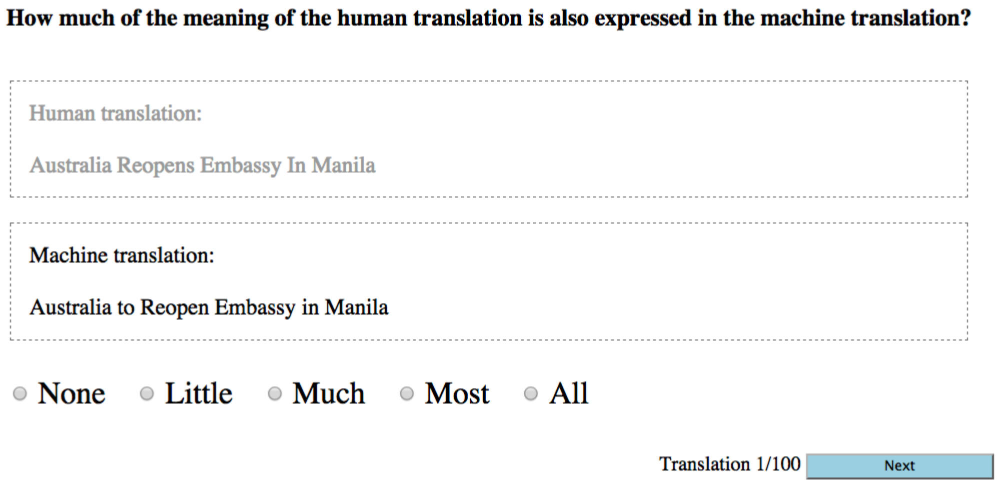
We note that even in such a scenario, human translations differ from each other. We measured the average similarity between the four references in the dataset using the Meteor evaluation metric (Denkowski and Lavie, 2014). Meteor scores range between 0 and 1 and reflect the proportion of similar words occurring in similar order. This metric is normally used to compare the MT output with a human reference, but it can also be applied to measure similarity between any two translations. We computed Meteor for all possible combinations between the four available references and took the average score. Even though Meteor covers certain amount of acceptable linguistic variation by allowing for synonym and paraphrase matching, the resulting score is only 0.33, which shows that, not surprisingly, human translations vary substantially.

To make the annotation process feasible given the resources available, we selected a subset of 100 source sentences for the experiment. To ensure variable levels of similarity between the MT and each of the references, we computed sentence-level Meteor scores for the MT outputs using each of the references and selected the sentences with the highest standard deviation between the scores.

3.2 Method

We developed a simple online interface to collect human judgments. Our evaluation task was based on the adequacy criterion. Specifically, judges were asked to estimate how much of the meaning of the human translation was expressed in the MT output (see Figure 1). The responses were interpreted on a five-point scale, with the labels in Figure 1 corresponding to numbers from 1 ("None") to 5 ("All").

For the main task, judgments were collected using English native speakers who volunteered to participate. They were either professional translators or researchers with a degree in Computational Linguistics, English or Translation Studies. 20 annotators participated in this monolingual task. Each of them evaluated the same set of 100 MT outputs. Our estimates showed that the task could be completed in approximately one hour. The annotators were divided into four groups, corresponding to the four available refer-



How much of the meaning of the human translation is also expressed in the machine translation?

Human translation:

Australia Reopens Embassy In Manila

Machine translation:

Australia to Reopen Embassy in Manila

Figure 1: Evaluation Interface

ences. Each group contained five annotators independently evaluating the same set of sentences. Having multiple annotators in each group allowed us to minimize the effect of individual annotators' biases, preferences and expectations.

As a control group, five annotators (native speakers of English, fluent in Chinese or bilingual speakers) performed a bilingual evaluation task for the same MT outputs. In the bilingual task, annotators were presented with an MT output and its corresponding source sentence and asked how much of the meaning of the source sentence was expressed in the MT.

In total, we collected 2,500 judgments. Both the data and the tool for collecting human judgments are available at `https://github.com/mfomicheva/tradopad.git`.

4 Reference Bias

The goal of the experiment is to show that depending on the reference translation used for evaluation, the quality of the same MT output will be perceived differently. However, we are aware that MT evaluation is a subjective task. Certain discrepancies between evaluation scores produced by different raters are expected simply because of their backgrounds, individual perceptions and expectations regarding translation quality.

To show that some differences are related to reference bias and not to the bias introduced by individual annotators, we compare the agreement between annotators evaluating with the same and with different references. First, we randomly select from the data 20 pairs of annotators who used the same reference translations and 20 pairs of annotators who used different reference translations. The agreement is then computed for each pair. Next, we calculate the average agreement for the same-reference and different-reference groups. We repeat the experiment 100 times and report the corresponding averages and confidence intervals.

Table 1 shows the results in terms of standard (Cohen, 1960) and linearly weighted (Cohen, 1968) Kappa coefficient (k).[1] We also report one-off version of weighted k, which discards the disagreements unless they are larger than one category.

Kappa	**Diff. ref.**	**Same ref.**	**Source**
Standard	.163±.01	.197±.01	0.190±.02
Weighted	.330±.01	.373±.01	0.336±.02
One-off	.597±.01	.662±.01	0.643±.02

Table 1: Inter-annotator agreement for different-references (Diff. ref.), same-reference (Same ref.) and source-based evaluation (Source)

As shown in Table 1, the agreement is consistently lower for annotators using different references. In other words, the same MT outputs systematically receive different scores when differ-

[1] In MT evaluation, agreement is usually computed using standard k both for ranking different translations and for scoring translations on an interval-level scale. We note, however, that weighted k is more appropriate for scoring, since it allows the use of weights to describe the closeness of the agreement between categories (Artstein and Poesio, 2008).

ent human translations are used for their evaluation. Here and in what follows, the differences between the results for the same-reference annotator group and different-reference annotator group were found to be statistically significant with p-value < 0.01.

The agreement between annotators using the source sentences is slightly lower than in the monolingual, same-reference scenario, but it is higher than in the case of the different-reference group. This may be an indication that reference-based evaluation is an easier task for annotators, perhaps because in this case they are not required to shift between languages. Nevertheless, the fact that given a different reference, the same MT outputs receive different scores, undermines the reliability of this type of evaluation.

	Human score	**BLEU score**
Reference 1	1.980	0.1649
Reference 2	2.342	0.1369
Reference 3	2.562	0.1680
Reference 4	2.740	0.1058

Table 2: Average human scores for the groups of annotators using different references and BLEU scores calculated with the corresponding references. Human scores range from 1 to 5, while BLEU scores range from 0 to 1.

In Table 2 we computed average evaluation scores for each group of annotators. Average scores vary considerably across groups of annotators. This shows that MT quality is perceived differently depending on the human translation used as gold-standard. For the sake of comparison, we also present the scores from the widely used automatic evaluation metric BLEU. Not surprisingly, BLEU scores are also strongly affected by the reference bias. Below we give an example of linguistic variation in professional human translations and its effect on reference-based MT evaluation.

Src: 不过这一切都由不得你[2]
MT: But all this is beyond the control of you.
R1: But all this is beyond your control.
R2: However, you cannot choose yourself.
R3: However, not everything is up to you to decide.
R4: But you can't choose that.

Although all the references carry the same message, the linguistic means used by the translators are very different. Most of these references are high-level paraphrases of what we would consider a close version of the source sentence. Annotators are expected to recognize meaning-preserving variation between the MT and any of the references. However, the average score for this sentence was 3.4 in case of Reference 1, and 2.0, 2.0 and 2.8 in case of the other three references, respectively, which illustrates the bias introduced by the reference translation.

5 Time Effect

It is well known that the reliability and consistency of human annotation tasks is affected by fatigue (Llorà et al., 2005). In this section we examine how this factor may gave influenced the evaluation on the impact of reference bias and thus the reliability of our experiment.

We measured inter-annotator agreement for the same-reference and different-reference annotators at different stages of the evaluation process. We divided the dataset in five sets of sentences based on the chronological order in which they were annotated (0-20, 20-40, ..., 80-100). For each slice of the data we repeated the procedure reported in Section 4. Figure 2 shows the results.

First, we note that the agreement is always higher in the case of same-reference annotators. Second, in the intermediate stages of the task we observe the highest inter-annotator agreement (sentences 20-40) and the smallest difference between the same-reference and different-reference annotators (sentences 40-60). This seems to indicate that the effect of reference bias is minimal half-way through the evaluation process. In other words, when the annotators are already acquainted with the task but not yet tired, they are able to better recognize meaning-preserving variation between different translation options.

To further investigate how fatigue affects the evaluation process, we tested the variability of human scores in different (chronological) slices of the data. We again divided the data in five sets of sentences and calculated standard deviation between the scores in each set. We repeated this procedure for each annotator and averaged the results. As can be seen in Figure 3, the variation between

[2] Literally: "However these all totally beyond the control of you."

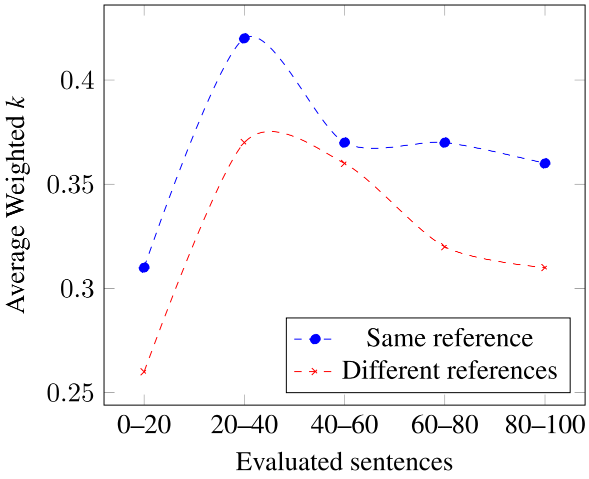


Figure 2: Inter-annotator agreement at different stages of evaluation process

the scores is lower in the last stages of the evaluation process. This could mean that towards the end of the task the annotators tend to indiscriminately give similar scores to any translation, making the evaluation less informative.

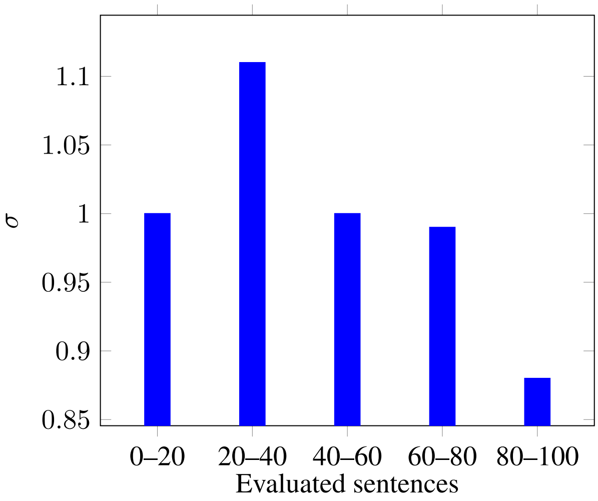


Figure 3: Average standard deviations between human scores for all annotators at different stages of evaluation process

6 Conclusions

In this work we examined the effect of reference bias on monolingual MT evaluation. We compared the agreement between the annotators who used the same human reference translation and those who used different reference translations. We were able to show that in addition to the inevitable bias introduced by different annotators, monolingual evaluation is systematically affected by the reference provided. Annotators consistently assign different scores to the same MT outputs when a different human translation is used as gold-standard. The MTs that are correct but happen to be different from a particular human translation are inadvertently penalized during evaluation.

We also analyzed the relation between reference bias and annotation at different times throughout the process. The results suggest that annotators are less influenced by specific translation choices present in the reference in the intermediate stages of the evaluation process, when they have already familiarized themselves with the task but are not yet fatigued by it. To reduce the fatigue effect, the task may be done in smaller batches over time. Regarding the lack of experience, annotators should receive previous training.

Quality assessment is instrumental in the development and deployment of MT systems. If evaluation is to be objective and informative, its purpose must be clearly defined. The same sentence can be translated in many different ways. Using a human reference as a proxy for the source sentence, we evaluate the similarity of the MT to a particular reference, which does not necessarily reflect how well the contents of the original is expressed in the MT or how suitable it is for a given purpose. Therefore, monolingual evaluation undermines the reliability of quality assessment. We recommend that unless the evaluation is aimed for a very specific translation task, where the number of possible translations is indeed limited, the assessment should be performed by comparing MT to the original text.

Acknowledgments

Marina Fomicheva was supported by funding from IULA (UPF) and the FI-DGR grant program of the Generalitat de Catalunya. Lucia Specia was supported by the QT21 project (H2020 No. 645452). The authors would also like to thank the three anonymous reviewers for their helpful comments and suggestions.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisboa, Portugal.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20:37–46.

Jacob Cohen. 1968. Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychological bulletin*, 70(4):213–220.

Trevor Cohn and Lucia Specia. 2013. Modelling Annotator Bias with Multi-task Gaussian Processes: An Application to Machine Translation Quality Estimation. In *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics*, pages 32–42.

Michael Denkowski and Alon Lavie. 2010. Choosing the Right Evaluation for Machine Translation: an Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks. In *Proceedings of the Ninth Biennal Conference of the Association for Machine Translation in the Americas*.

Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, pages 376–380.

Markus Dreyer and Daniel Marcu. 2012. HyTER: Meaning-equivalent Semantics for Translation Evaluation. In *Proceedings of 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–171.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41.

Francisco Guzmán, Ahmed Abdelali, Irina Temnikova, Hassan Sajjad, and Stephan Vogel. 2015. How do Humans Evaluate Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 457–466.

Xavier Llorà, Kumara Sastry, David E Goldberg, Abhimanyu Gupta, and Lalitha Lakshmi. 2005. Combating User Fatigue in iGAs: Partial Ordering, Support Vector Machines, and Synthetic Fitness. In *Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation*, pages 1363–1370.

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 of the ACL*, pages 311–318.

Mark Przybocki, Kay Peterson, and Sebastian Bronsart. 2008. Official Results of the NIST 2008 "Metrics for MAchine TRanslation" Challenge (MetricsMATR08). In *Proceedings of the AMTA-2008 Workshop on Metrics for Machine Translation*, Honolulu, Hawaii, USA.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine Translation Evaluation versus Quality Estimation. *Machine Translation*, 24(1):39–50.

John White, Theresa O'Connell, and Francis O'Mara. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. In *Proceedings of the Association for Machine Translation in the Americas Conference*, pages 193–205, Columbia, Maryland, USA.