A fuzzier approach to machine translation evaluation: A pilot study on post-editing productivity and automated metrics in commercial settings

Carla Parra Escartín
Hermes Traducciones
C/ Cólquide 6, portal 2, 3.º I
28230 Las Rozas, Madrid, Spain
carla.parra@hermestrans.com

Manuel Arcedillo
Hermes Traducciones
C/ Cólquide 6, portal 2, 3.º I
28230 Las Rozas, Madrid, Spain
manuel.arcedillo@hermestrans.com

Abstract

Machine Translation (MT) quality is typically assessed using automatic evaluation metrics such as BLEU and TER. Despite being generally used in the industry for evaluating the usefulness of Translation Memory (TM) matches based on text similarity, fuzzy match values are not as widely used for this purpose in MT evaluation. We designed an experiment to test if this fuzzy score applied to MT output stands up against traditional methods of MT evaluation. The results obtained seem to confirm that this metric performs at least as well as traditional methods for MT evaluation.

1 Introduction

In recent years, Machine Translation Post-Editing (MTPE) has been introduced in real translation workflows as part of the production process. MTPE is used to reduce production costs and increase the productivity of professional translators. This productivity gain is usually reflected in translation rate discounts. However, the question of how to assess Machine Translation (MT) output in order to determine a fair compensation for the post-editor is still open.

Shortcomings of traditional metrics, such as BLEU (Papineni et al., 2001) and TER (Snover et al., 2006), when applied to MTPE include unclear correlation with productivity gains, technical difficulties for their estimation by general users and lack of intuitiveness. A more common metric already used in translation tasks for evaluating text similarity is the Translation Memory (TM) fuzzy match score. Based on the fuzzy score analysis, rate discounts due to TM leverage are then applied. We designed an experiment to test if this fuzzy score applied to MT output stands up against traditional methods of MT evaluation.

The remainder of this paper is structured as follows: Section 2 presents the rationale behind the experiment. Section 3 explains the pilot experiment itself. Section 4 reports the results obtained and what they have revealed, and finally Section 5 summarizes our work and discusses possible paths to explore in the light of our findings.

2 Rationale

As far as MT evaluation is concerned, a well-established evaluation metric is BLEU, although it has also received criticism (Koehn, 2010). It is usually considered that BLEU scores above 30 reflect understandable translations, while scores over 50 are considered good and fluent translations (Lavie, 2010). However, the usefulness of "understandable" translations for MTPE is questionable. Contrary to other MT applications, post-editors do not depend on MT to understand the meaning of a foreign-language sentence. Instead, they expect to re-use the largest possible text chunks to meet their client's requirements, regardless of the meaning or fluency conveyed by the raw MT output. This criticism also holds true for human annotations on Adequacy and Fluency[1].

Other metrics more focused in post-editing effort have been developed, such as TER. However, how should one interpret an improvement in BLEU score from 45 to 50 in terms of productivity? Likewise, does a TER value of 35 deserve any kind of discount? Most likely, the vast majority of translators would

[1]For details of these scores see, for example, the TAUS adequacy and fluency guidelines at https://www.taus.net/.

Proceedings of the ACL 2015 Fourth Workshop on Hybrid Approaches to Translation (HyTra), pages 40–45,
Beijing, China, July 31, 2015.

be unable to answer these questions and yet they would probably instantly acknowledge that fuzzy text similarities of 60% are not worth editing, while they would be happy to accept discounts for 80% fuzzy scores based on an analogy with TM matches. Organizations such as TAUS have already proposed alternative models which use fuzzy matches for MT evaluation, such as the "MT Reversed analysis".[2]

In order to compare alternative measures based on fuzzy matches with BLEU and TER scores, we designed an experiment involving both MTPE and translating from scratch in a real-life translation scenario. It is worth noting that the exact algorithm used by each Computer Assisted Translation (CAT) tool for computing the fuzzy score is unknown[3]. In this paper, we use the Sørensen-Dice coefficient (Sørensen, 1948; Dice, 1945) for fuzzy match scores, unless otherwise specified.

3 Pilot experiment settings

Following similar works (Federico et al., 2012), the experiment aimed to replicate a real production environment. Two in-house translators were asked to translate the same file from English into Spanish using one of their most common translation tools (memoQ[4]). This tool was chosen because of its feature for recording time spent in each segment. Other tools which also record this value and other useful segment-level indicators, such as keystrokes[5], or MTPE effort[6], were discarded due to them not being part of the everyday resources of the translators involved in the experiment. Translators were only allowed to use the TM, the terminology database and the MT output included in the translation package. Other memoQ's productivity enhancing features were disabled (especially, predictive text, sub-segment leverage and automatic fixing of fuzzy matches) to allow better comparisons with translation environments which may not offer similar features.

3.1 Text selection

The file to be translated had to meet the following requirements:

1. Belong to a real translation request.
2. Originate from a client for which our company owned a customized MT engine.
3. Have a word volume capable of engaging translators for several hours.
4. Include significant word counts for each TM match band (i.e., exact matches, fuzzy matches and no-match segments[7]).

The original source text selected contained over 8,000 words and was part of a software user guide. All repetitions and internal leverage segments were filtered out to avoid skewing due to the inferior typing and cognitive effort required to translate the second of two similar segments. During this text selection phase, we studied the word counts available for all past projects of this client, which were already generated using a different tool (SDL Trados Studio[8]) than the one finally used in the experiment (memoQ). Table 1 shows the word counts of our text according to both tools.

	memoQ		Trados Studio	
TM match	**Words**	**Seg.**	**Words**	**Seg.**
100%	1226	94	1243	95
95-99%	231	21	1044	55
85-94%	1062	48	747	43
75-84%	696	42	608	42
No Match	3804	263	3388	233
Total	**7019**	**468**	**7030**	**468**

Table 1: Final word counts.

As Table 1 shows, CAT tools may differ greatly in the word counts and fuzzy match distribution. As Studio showed significant word volumes for every band, the file used for the test seemed appropriate. However, when using memoQ one of the fuzzy match bands (95-99%) ended up with significantly less words than the other bands. At the same time, there was an increase in no-match segments. This provided a more solid sample

[2]See the pricing MTPE guidelines at `https://www.taus.net`.

[3]It is believed that most are based on some adjustment of Levenshtein's edit distance (Levenshtein, 1965).

[4]The version used was memoQ 2015 build 3.

[5]For example, PET (Aziz et al., 2012) and iOmegaT (Moran et al., 2014).

[6]For example, MateCat (Federico et al., 2012).

[7]In general, any TM fuzzy match below 75% is considered a no-match segment due to the general acceptance that such leverage does not yield any productivity increase.

[8]The version used was SDL Trados Studio 2014 SP1.

for comparing MTPE and translation throughputs, increasing the count to 3804 words, half of which were randomly selected for MTPE using the test set generator included in the m4loc package[9]. Table 2 shows word counts after this division.

Origin	Words	Segments
No Match (MTPE)	1890	131
No Match (Translation)	1914	132
Total	**3804**	**468**

Table 2: No-match word count distribution after random division.

3.2 MT engine

The system used to generate the MT output was Systran's[10] RBMT engine. This is the system normally used in our company for post-editing machine translated texts from this client. It can be considered a mature engine, since at the time of the experiment it had been subject to ongoing in-house customization for over three years via dictionary entries, software settings, and pre- and post-editing scripts, as well as having a consistent record for productivity enhancement. Although Systran includes a Statistical Machine Translation (SMT) component, this was not used in our experiment because in previous tests it produced a less adequate MT output for MTPE.

3.3 Human translators

Both translators involved had five years' experience in translation. However, Translator 2 also had three years' experience in MTPE and had been involved in previous projects of this client. Translator 1 did not have any experience either in MTPE or with the client's texts. They were assigned a hand-off package which included all necessary files and settings for the experiment. They were asked to translate the file included in the package performing all necessary edits in the MT output and TM matches to achieve the standard full quality expected by the client.

4 Results and discussion

Once the translation and MTPE task was delivered by both translators, we analyzed their output using different metrics:

1. **Words per hour**: Amount of words translated/post-edited per hour, according to memoQ's word count and time tracking feature.
2. **Fuzzy match**: Based on the Sørensen-Dice coefficient, this metric is a statistic used to compare the similarity of two samples. We used the Okapi Rainbow library[11]. The comparison is based in 3-grams.
3. **BLEU**: Widely used for MT evaluation. It relies on n-gram overlapping.
4. **TER**: Another widely used metric, based on the number of edits required to make the MT output match a reference.
5. **Productivity gain**: Based on the number of words translated/post-edited per hour, we estimated the productivity gain for each band when compared to unaided translation throughput.

For the metrics involving a comparison, we compared the TM match suggestion or MT raw output against the final delivered text by the translators. The results of our evaluation are reported in Table 3.

	W/h	Fuzzy	BLEU	TER	Prod. gain %
Trans.1					
100%	1542	97.50	91.96	4.64	65.20
95-99%	963	92.43	87.91	6.91	3.14
85-94%	1158	90.92	80.19	13.02	24.12
75-84%	1120	87.93	73.94	19.08	20.03
PE	910	88.53	69.57	18.89	-2.46
TRA	933	-	-	-	-
Trans. 2					
100%	2923	97.91	92.38	3.99	121.61
95-99%	2625	92.76	89.35	6.37	99.05
85-94%	2237	91.19	81.00	12.69	69.61
75-84%	1585	85.21	71.98	21.03	20.17
PE	1728	87.74	66.37	20.98	31.00
TRA	1319	-	-	-	-

Table 3: Results obtained for both translators.

Both translators had unusually high throughputs for MTPE and unaided translation, especially when compared to the standard reference of 313-375 words per hour (2500-3000 words per day). Taking this as reference, Translator 1 would have experienced more than 140% productivity increase, while Translator 2 would have translated at least 350% faster. However, despite this high MTPE speed, Translator 1 did not experience

[9] https://code.google.com/p/m4loc/

[10] Systran 7 Premium Translator was used. No language model was applied.

[11] http://okapi.opentag.com/

any productivity gain (quite the contrary), while Translator 2 saw a productivity increase of "just" 31%. This may point out that the faster texts to translate are also the fastest to post-edit. Thus the importance of having an unaided translation reference for each sample instead of relying on standard values (Federico et al., 2012).

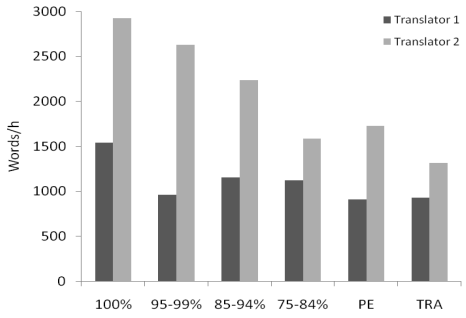

Figure 1: Productivity in words per hour of both translators.

The difference between the MT benefit for both translators might be due to the little MTPE experience of Translator 1. Furthermore, Translator 2 was already familiarized with the texts of this client, while it was the first time Translator 1 worked with them. Presumably, Translator 1 had to spend more time acquiring the client's preferred terminology and performing TM concordance lookups to achieve consistency with previously translated content. This seems to have negated part of the benefits of fuzzy matching and MT output leverage (see the flatness of Translator 1's throughputs for fuzzy, MTPE and translation bands in Figure 1). Translator 2 does show a distinct throughput for each category.

Another possible explanation for Translator 1's performance would be that the quality of the raw MT output is low. However, Translator 2's productivity gains and comparison with past projects' performance contradict this. We therefore concluded that the most probable explanation to the difference in terms of productivity might be due to the MTPE experience of both translators. In fact, studies about impact of translator's experience agree that more experienced translators do MTPE faster (Guerberof Arenas, 2009), although they do not usually distinguish between experience in translation and experience in MTPE.

Figures 2 and 3 plot the productivity for each band against the different evaluation measures discussed for Translator 1 and 2, respectively. TER has been inverted for a more direct comparison.

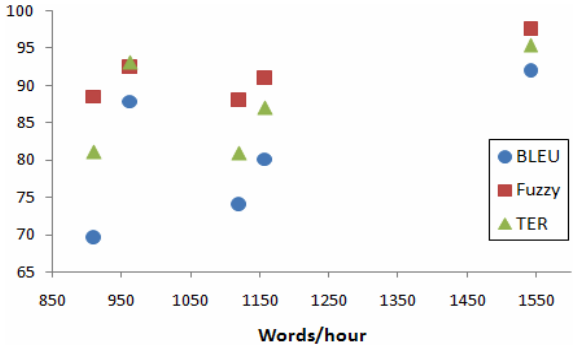

Figure 2: Productivity vs. automated measures for Translator 1.

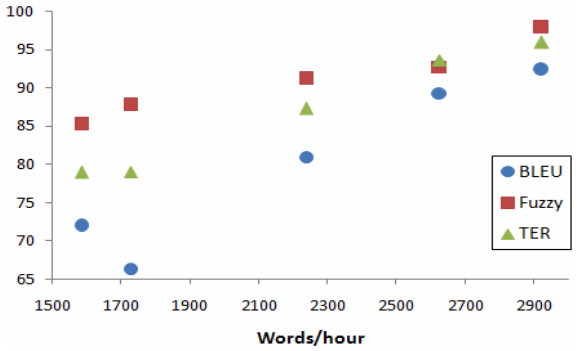

Figure 3: Productivity vs. automated measures for Translator 2.

It is remarkable that Translator 2's MTPE throughput was even higher than the one for the lowest fuzzy match band. According to BLEU (66.37 vs. 71.98), the situation should have been the opposite, while according to TER both throughputs should have been more or less the same (20.98 vs. 21.03). The fuzzy match value (87.74 vs. 85.21) is the only one from the chosen set of metrics to reflect the higher throughput of the MTPE sample over the 75-84% band.

Despite this fact, all three metrics showed a strong correlation with productivity for Translator 2, while the fuzzy score had the strongest correlation for Translator 1 (see Table 4). Based on the results obtained, the fuzzy score could be used in MTPE scenarios as a valid alternative metric for evaluating MT output.

Despite not being used often in research, we have found out that it could give a good insight on the translation quality of MT output, as it performs as good as or even better than the other metrics evaluated. At the same time, as it is a well-established metric in translation business, it might be easier for translators to understand and assess MTPE tasks.

	r_{fuzzy}	r_{BLEU}	r_{TER}
Trans. 1	0.785	0.639	0.568
Trans. 2	0.975	0.960	0.993

Table 4: Pearson correlation between productivity and evaluation measures.

Finally, another advantage of the fuzzy metric is the fact that it does not depend on tokenization. It is a well known-fact that depending on the tokenization applied to the MT output and the reference, differences in BLEU arise. This is illustrated in Table 5, which reports the BLEU scores obtained for the MT post-edited text as estimated by different tools: Asiya (Giménez and Màrquez, 2010), Asia Online's Language Studio[12], and the multibleu script included in the SMT system MOSES (Koehn et al., 2007). As can be observed, there are significant differences in BLEU scores for the same band for both translators.

	BLEU (Asiya)	**BLEU (Asia Online)**	**BLEU (MOSES)**
Trans. 1			
100%	91.96	91.94	91.96
95-99%	87.91	87.77	87.20
85-94%	80.19	80.12	80.20
75-84%	73.94	74.27	74.09
PE	69.57	68.93	69.37
Trans. 2			
100%	92.38	92.37	92.39
95-99%	89.35	89.22	89.19
85-94%	81.00	80.94	80.97
75-84%	71.98	72.55	72.12
PE	66.37	65.66	66.16

Table 5: BLEU results as computed by different evaluation tools.

5 Conclusion and Future work

In this paper, we have reported a pilot experiment based on a real-life translation project. The translation job was analyzed with the usual CAT tools in our company to ensure the project included samples of all TM match bands. All matches below 75% TM fuzzy leverage were then split into two parts: one was used for MTPE, and the other half was translated from scratch. The raw MT output was generated by a customized Systran RBMT system and integrated in the CAT environment used to run the experiment.

We have discovered that MT quality may also be assessed using a fuzzy score mirroring TM leverage (we used 3-gram Sørensen-Dice coefficient). It correlates with productivity as well as or even better than BLEU and TER, it is easier to estimate[13], and does not depend on tokenization. Moreover, this metric is more familiar to all parties in the translation industry, as they already work with fuzzy matches when processing translation jobs via CAT tools.

Another interesting finding is that MTPE might result in an increased productivity ratio if the translator already has MTPE experience and is familiarized with the client's texts. However, further research on this matter is needed to confirm the impact of each factor separately.

The results of this pilot study reveal that a "fuzzier" approach might be a valid MTPE evaluation measure. In future work we plan to repeat the experiment with more translators to see if the findings reported here replicate. We believe that the proposed fuzzy-match approach, if proven valid, would be more easily embraced in MTPE workflows than more traditional evaluation measures.

Acknowledgments

The research reported in this paper is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement nº 317471.

References

Wilker Aziz, Sheila C. M. de Sousa, and Lucia Specia. 2012. PET; a Tool for Post-Editing and Assessing Machine Translation. In *Eighth International Conference on Language Resources and Evaluation (LREC 12)*, pages 3982–3987, Istanbul, Turkey, May. ELRA.

Lee R. Dice. 1945. Measures of the Amount of

[12] `http://www.asiaonline.net/EN/Default.aspx`

[13] There are free open-source tools (e.g. Okapi Rainbow) which support industry-standard bilingual files (XLIFF and TMX) able to calculate fuzzy scores.

Ecologic Association Between Species. *Ecological Society of America*, 26(3):297–302, July.

Marcello Federico, Alessandro Cattelan, and Marco Trombetti. 2012. Measuring User Productivity in Machine Translation Enhanced Computer Assisted Translation. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*, San Diego, CA, October. AMTA.

Jesús Giménez and Lluís Màrquez. 2010. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86.

Ana Guerberof Arenas. 2009. Productivity and quality in the post-editing of outputs from translation memories and machine translation. *The International Journal of Localisation*, 7, Issue 1:11–21.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic, June. ACL.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

Alon Lavie, 2010. *Evaluating the Output of Machine Translation Systems*. AMTA, Denver, Colorado, USA, October.

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertionsm, and reversals. *Soviet Physics Doklady*, 10(8):707–710, February.

John Moran, Christian Saam, and Dave Lewis. 2014. Towards desktop-based CAT tool instrumentation. In *Proceedings of the Third Workshop on Post-Editing Technology and Practice*, pages 99–112, Vancouver, BC, October. AMTA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, September.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA, August.

Thorvald Sørensen. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab*, 5 (4):1–34.