The Trials and Tribulations of Predicting Machine Translation Post-Editing Productivity

Lena Marg
Welocalize, Inc.
Frederick, MD, United States
E-mail: lena.marg@welocalize.com

Abstract

While an increasing number of (automatic) metrics is available to assess the linguistic quality of machine translations, their interpretation remains cryptic to many users, specifically in the translation community. They are clearly useful for indicating certain overarching trends, but say little about actual improvements for translation buyers or post-editors. However, these metrics are commonly referenced when discussing pricing and models, both with translation buyers and service providers. With the aim of focusing on automatic metrics that are easier to understand for non-research users, we identified Edit Distance (or Post-Edit Distance) as a good fit. While Edit Distance as such does not express cognitive effort or time spent editing machine translation suggestions, we found that it correlates strongly with the productivity tests we performed, for various language pairs and domains.

This paper aims to analyse Edit Distance and productivity data on a segment level based on data gathered over some years. Drawing from these findings, we want to then explore how Edit Distance could help in predicting productivity on new content. Some further analysis is proposed, with findings to be presented at the conference.

Keywords: Machine Translation, Metrics, Quality Evaluation, Edit Distance, Post-Edit Productivity, Predicting Productivity

1. Introduction

There is an increasing number of metrics available to assess the linguistic quality produced by machine translation (MT) systems. Using several of them in parallel helps to validate the individual scores and thus increase confidence in the results. However, while they are very useful for indicating quality trends (for example system B with a BLEU of 79 can quite safely be expected to be better than old system A with a BLEU of 60), their interpretation remains cryptic to many users: "What does it mean if my English to Polish MT system gets a BLEU of 50, is this good or bad?"; "Does a GTM of 89 tell me that the translations are correct and reliable in 89% of cases?"

These questions are of particular interest and relevance when the machine translations are used for very specific purposes, such as to enable critical, real-time communication, or when they are intended for human post-editing.

There appears to also be a certain divide, whereby non-linguists prefer automatic scores as they are deemed mathematically more reliable and less subjective, while translators put more trust into "visual" quality checks, such as scoring or reviewing samples themselves.

The metric that seems to nicely bridge the gap then would be Edit Distance (or the Levenshtein algorithm)[1], in so far as it is an algorithm and outputs a score, and yet allows translators to review and understand it on an actual side-by-side comparison. It is also attractive with translation buyers, who will often be overwhelmed by a plethora of seemingly black box scores.

The objective of the analysis proposed in this paper is to take a deeper look at different expressions of Edit Distance at a segment-level, and how reliable they are as a quality metric for estimating post-editing effort and possible productivity gains.

2. Contextualization

In our work as a Language Service Provider (LSP), providing translation services into a wide range of languages, clients increasingly approach us with requests for machine translation solutions to reduce translation cost or turnaround time. While there is an increased demand for raw machine translation, the diversification of content and an uptake of dynamic quality-models simultaneously lead to an increase in the demand for different levels of human post-editing.

The integration of machine translation into a program always requires some form of measurement of the quality of the raw machine translation, both as part of the MT system customisation as well as ensuring the required quality needed for the specific purpose (publishing MT un-edited or with different levels of human post-editing) is met.

At Welocalize, the standard approach for evaluating the quality of raw MT output covers a range of automatic metrics (BLEU, GTM, TER, Nist, Meteor, Edit Distance, Precision and Recall) as well as a human assessment typically performed by a linguist. This human evaluation

[1] See for example http://www.levenshtein.net/

can either consist of scoring translation units in a sample on a 1-5 scale for accuracy and fluency or utility, error typology, or a so-called productivity test. Productivity tests come into play when the machine translation output is intended to be post-edited to an agreed level of quality by a linguist / translator / post-editor. They are typically performed on a sample equivalent to one or two standard days of translation work using a tool such as iOmegat[2], MateCat or the Qualitivity plug-in developed for SDL Studio, which measure time spent in each translation segment, number of segment visits, key strokes and so forth. They can either be set-up to simply measure the total time spent on a post-editing test, or a more elaborate approach is to provide the tester with some MT suggestions and some empty segments that need to be translated from scratch. In the latter scenario, we obtain a productivity delta, in other words a percentage indicating how much faster or slower the tester was using machine translation over translating "from scratch".

Automatic metrics are typically based on some notion of comparing the machine translation proposals to some previously obtained "gold standard" translation of the same source text[3]. Human assessment can do without a pre-existent reference translation, and the purpose is usually newly defined with each evaluation ask, for instance "utility" of the MT proposals to potential end-users, "accuracy" and "fluency" for end-users or different levels of post-editing, "productivity" again with an eye to post-editing, side-by-side ranking of different machine translation proposals, error typology and so on and forth. Both categories of metrics have their advantages and disadvantages in terms of criteria such as sample selection, reliability, obtaining data (reference translations), time and cost involved for running them etc.

Primarily due to the effort and cost required to run human evaluations, there always remains, however, an interest in moving to an entirely automated approach.

Two questions then remain of key interest to us:

- Can the automatic metrics reliably replace the human verdict on machine translation quality?
- Can we reliably predict post-editor productivity gains purely from a range of automatic metrics obtained on the raw machine translation?

During the MTE workshop at the LREC 2014 conference, I presented the findings of a study correlating a number of metrics, based on data collated in our internal database of past machine translation evaluations. One finding was that the automatic metrics investigated seemed to correlate well with each other, and so did the human metrics, however correlation between BLEU and human quality scoring, for example, was weak.

3. Research Proposal

We propose to undertake an analysis in two phases:

3.1 Phase 1 - Validation of existing correlations

Revisit the correlations, for a number of language pairs to validate our original findings. Our database has grown since the last analysis and this will hopefully provide further insights.

3.2 Phase 2 - Edit Distance and Productivity on the segment-level

Analyse and measure correlations on a subset of metrics, at segment-level, that we think are most likely to help us make predictions about post-editor productivity for future machine translation programs. As outlined above, these are Edit Distance score (Levenshtein), Edit Distance expressed as percentage and post-editing productivity (time spent in each segment plus segment visits).

This phase represents the core of the analysis to be undertaken.

3.2.2. Why look at the segment level?

Segment-level analysis is interesting for a number of reasons:

- It can provide further insights as to what type of segments (for example short versus long) really work best with machine translation and are therefore most beneficial for post-editors;
- It will allow a closer look at what the Levenshtein Score / Edit Distance really mean from a post-editing perspective (does a 20% Edit Distance on a single word represent the same effort as a 20% Edit Distance on a 25-word sentence for a post-editor?);
- A segment-level analysis would allow for easier comparison with the segment-level fuzzy score assigned to Translation Memory matches (100%, Fuzzy), typically deployed in post-editing projects alongside machine translation. This is furthermore relevant with new technologies gaining in interest that would pre-select the "best" candidate from a number of machine translation and / or Translation Memory matches for the given translation segment.

Thanks to frequent evaluations on our growing list of MT programs, we have access to a significant database of completed evaluations covering some or all of the above metrics, various content types and language pairs.

Using a range of proprietary technologies, we can produce summary reports showing productivity, Edit Distance scores and percentages in an aligned and easy-to-analyse format.

[2] An instrumented version of the open source translation tool OmegaT, created in collaboration between Welocalize, John Moran, and the Centre for Next Generation Localization (CNGL).

[3] In an ideal scenario, several "gold standard" references would be used, but this is difficult to achieve in a commercial setting, where a translation will only be requested and paid for once.

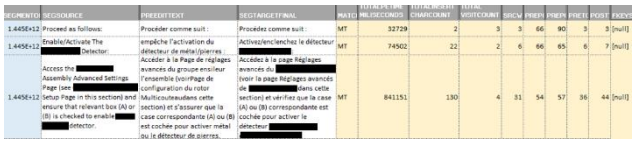

Figure 1: Segment-level PE Distance and productivity information.

In order to narrow down the scope of analysis, the language pairs and content type with most available data will be selected and limited to a maximum of three pairs.
The data samples will all be taken from completed productivity tests performed using one of the technologies mentioned above, looking at time-spent on a segment-level, segment visits and keystrokes performed.

4. The Metrics: Edit Distance and Post-Editing Productivity

4.1 Edit Distance - Score

In our context, the Levenshtein algorithm / Edit Distance is commonly used to measure how many (character) edits are required to transform a given string of raw MT output to the final, *post-edited* version of the same string that is compliant with the translation quality requirements specified by the buyer. It provides a segment-level score. However, the Edit Distance score by itself does not take into account the actual time spent on a given translation unit and provides an absolute score, whereby the maximum number of edits possible per segment would be the total number of characters of the reference translation. In other words, if the "gold standard" reference translation has 25 characters, the Edit Distance Score can be 25 at most.

4.2 Edit Distance - Percentage

In addition to the Edit Distance score, the PE Distance *percentage* adds the ratio of edits compared to the length of the final translation. This percentage arguably expresses the actual effort for the post-editor more accurately. This percentage is also frequently quoted in the context of the post-editor recruitment process or as explanation to translation buyers, as MT output, reference translation and expected effort are shown side-by-side.
However, it still does not consider the actual time spent making the edits. Looking at individual examples, it is not entirely obvious how reliably post-editing productivity gains can be deducted.

4.2.1 Example

To illustrate with an example:

- Two translation units, one with two words, one with ten.
- In the two-word unit, two character were replaced by one, to meet the new German spelling requirements (spelling).
- In the ten-word unit, a three-digit number had to be replaced with a two-digit number, to match the units in the source text (accuracy).

The Edit Distance score shows that the two-word segment required two edits, while the ten-word segment required three edits to be transformed – more, but not significantly so.
The Edit Distance percentage in turn suggests that the effort for the longer segment was significantly lower, as only as only 3.26% of the raw MT output need to be corrected.
From a post-editing perspective, however, the ten-word segment will take longer to correct, as there is more information to process and spotting a numerical error is typically harder, and yet also more critical in terms of the accuracy of the final translation.

Source	Hypothesis	Reference	Lev. Di	PE Dist. (% of ref. le	Wor
Stop Jerk	Stoss stoppen	Stoß stoppen	2	16.67%	2
Valid values are:[x1]0.005 to 25.0 mm/sec [x2]Single precision floating format	Gültige Werte sind:[x1]0,005 bis 500,0 mm/Sek.[x2] Gleitkommaformat mit einfacher Genauigkeit	Gültige Werte sind:[x1]0,005 bis 25,0 mm/Sek.[x2] Gleitkommaformat mit einfacher Genauigkeit	3	3.26%	10

Figure 2: Edit Distance score versus percentage / string length.

4.3 Post-Editing Productivity Gain

In an LSP context, it is common to assess the quality of machine translation for use by post-editors through so-called productivity tests[4]. The standard setup is described above under 2. The information obtained, on a segment-level provides further valuable information on the quality of the MT candidates and the post-editing effort.

SEGSOURCE	PREEDITTEXT	SEGTARGETFINAL	MA	TOTAL PE TIME MILISECOND	TOTAL INSERT CHAR COUNT	TOTAL VISIT COUNT	SRC WORD COUNT
Access the Video here:	Accedere a video qui:	È possibile accedere al video qui:	MT	101235	15	2	4
Access the Video here: (AVAILABLE IN ENGLISH ONLY)	Accedere a video qui: (DISPONIBILITÀ IN INGLESE SOLO)	È possibile accedere al video qui: (DISPONIBILE SOLO IN INGLESE)	MT	27172	18	1	8
According to the National Institute of Standards and Technology (NIST)1, three of the most common cloud deployment models are:	Secondo (National Institute of Standards and Technology) NIST il 1, tre del cloud più comuni sono:	Secondo il National Institute of Standards and Technology (NIST) 1, i tre modelli di implementazione del cloud più diffusi sono i seguenti:	MT	84344	60	1	20
Add fabric interconnects to your solution:	Aggiungere i Fabric Interconnect alla propria soluzione:	Aggiungere un fabric Interconnect alla propria soluzione:	MT	69734	5	3	6

Figure 3: productivity analysis showing source, MT, post-edited text with milliseconds spent, number of changes made, segment visit times and source character count.

5. Conclusion

Initial correlations between Edit Distance and productivity on a document-level seemed to suggest a strong correlation between these two metrics. However, both are heavily dependent on the post-editing behaviour of the individual translator. We therefore plan to revisit correlations established in the past with newly added data, to see if the initial trend is confirmed. In addition, a subset of languages and tests is to be analysed at a segment level, to allow for further conclusions on how useful Edit Distance might be

[4] Far more common than other approaches for measuring cognitive effort such as eye-tracking. By analysing the time spent in each segment to post-edit, the number of visits, and the changes made, we can still draw conclusions on specific parts within a segment that lead to the highest cognitive effort (e.g. harmonizing terminology in the very end, or converting measurements).

in predicting productivity gains, assisting with pre-selecting segments suitable for post-editing / not, and thereby aide in the discussion of pricing models.
Results will be shared as part of the presentation at the conference.

6. Acknowledgement

I would like to thank my colleagues Elaine O'Curran, Naoko Miyazaki, David Clarke and David Landan for their contribution.

7. Bibliography

O'Brien, S. (2011) Towards predicting post-editing productivity. Machine Translation, 25 (1). pp. 197-215. http://doras.dcu.ie/17154/1/Towards_Predicting_Postediting_Productivity_Final_2.pdf.
Accessed 17 March 2016

Papineni et al. (2002) BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.
http://www.aclweb.org/anthology/P02-1040.pdf
Accessed 17 March 2016

Specia, L. and Farzindar, A. (2010) Estimating Machine Translation Post-Editing Effort with HTER. AMTA-2010 Workshop Bringing MT to the User: MT Research and the Translation Industry. Denver, Colorado. http://clg.wlv.ac.uk/papers/show_paper.php?ID=279.
Accessed 17 March 2016

Tatsumi, M. (2009) Correlation between Automatic Evaluation Metric Scores, Post-Editing Speed, and Some Other Factors. MT Summit XII: proceedings of the twelfth Machine Translation Summit, pages 332-339 August 26-30, 2009, Ottawa, Ontario, Canada.
http://mt-archive.info/MTS-2009-Tatsumi.pdf.