Assumptions, Expectations and Outliers in Post-Editing

Laura Casanellas
Welocalize / Dublin, Ireland
laura.casanellas@welocalize.com

Lena Marg
Welocalize / London
lena.marg@welocalize.com

Abstract

As a multilingual vendor, we have access to machine translation (MT) scoring and other evaluation data on a wide range of language combinations and content types; we also have experience with different MT systems in production. Our daily work involves the collaboration with a wide spectrum of translation partners, from very MT-savvy to novices in this area. Being exposed to MT in such a varied and large-scale setup, we would like to share some of our insights into assumptions, expectations and outliers observed with regard to MT quality, productivity and suitability with a particular focus on the challenges that (individual) post-editor behavior presents in this context. Our observations are based on data correlations carried out at the end of 2013 from a database that contains all evaluation data produced during this year, as well as recent surveys with some of our very MT-savvy translation partners for deeper, locale-specific insights.

1 Introduction

In our company machine translation (MT) is typically integrated in the translation workflow as a productivity tool complementing translation memories, glossaries etc., with translators carrying out the required levels of post-editing. Content is translated into a multitude of languages (mostly from English) and MT is currently being used in production on a wide range of content types, from technical communication, user interface and corporate communication to user-generated contents. Additionally, we do not work with a specific MT system, but rather a variety of MT systems are evaluated and used – based on our own or our client's recommendations. At the end of 2013, we created a comprehensive database of results from automatic and human scorings of MT output as well as results from productivity tests obtained in that year, covering all these variables (locales, content types, MT systems).

Our productivity test shows the potential productivity gains obtained by moving from the task of translating to post-editing.

While the analysis and correlations drawn from this database confirmed certain assumptions, it allowed us to reassess expectations and also provided insights into outliers. In this abstract and our presentation, we would like to discuss these assumptions, expectations and outliers, benefitting from the wide range of variables used in the company. In this context, we want to draw attention to individual translator behavior, which might need to be considered more strongly when assessing MT output quality and usability.

2 The Database

The database mentioned above was created with all available data related to MT evaluation from 2013. The timeframe was delimited to one year.

Objectives for creating the database were multiple, but a key aim was to see if a correlation of currently available, internal data would help us make productivity predictions on future MT post-editing effort with the metrics currently in use in the company.

The categories included in the 2013 database are: client name, content type, locale, translation partner carrying out any human evaluations, BLEU, PE distance, human adequacy & fluency scores, productivity test deltas (in percent), productivity test throughputs (words post-edited versus words translated), MT system provider,

© 2014 by Welocalize.

owner of MT system maintenance (e.g. client, provider or Welocalize), comment on whether the test resource had received training on MT and PE, final quality scores (i.e.: the final translated / post-edited product).

2.1 Data Correlations

After populating the database with data on all the above categories, we started looking into correlations between different variables, e.g.: Adequacy versus BLEU, Fluency versus PE distance, Adequacy versus productivity delta, Fluency versus productivity delta, etc., using ***Pearson's r***. At this stage, we *intentionally* tried to keep the data sets broad, e.g. include all locales that had partaken in a given productivity test, rather than limit it to a few; include a range of MT systems rather than focusing only on one; including all post-editor profiles, rather than distinguishing between experienced and novice. To some extend the idea was precisely not to start with assumptions from the outset (like "engine X will anyway perform better than engine Y for Russian", "your translators are more open-minded to MT and will perform better" etc.). We wanted to see whether trends would emerge at a high level - trends that could be useful for us to dig into deeper in future or to exploit more with regard to productivity predictions for instance. This approach is further justified by the fact that our MT programs tend to cover various languages and content types and MT systems are often defined by the client, who would typically only invest into one MT system, unless this system offers only limited language pairs. In other words: MT systems are not only chosen on the basis of what the general assumption of their performance is, but also for cost and maintenance reasons.

Some assumptions were certainly confirmed by the data correlation. For instance the Adequacy score proved to be more strongly correlated to productivity deltas and the Fluency score to PE distance.

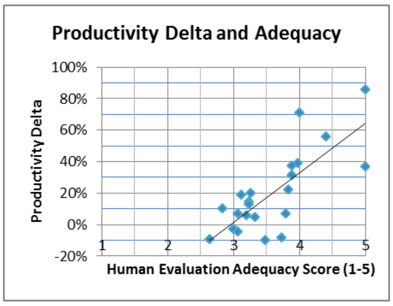

Fig 1 Productivity and Adequacy across all locales with a cumulative Pearson's r of 0.71, a very strong correlation

We find these correlations meaningful, as the final productivity tests are measured against our standard Quality Metrics and requirements for the respective content. For example, if Fluency scores and productivity delta do not correlate strongly, this suggests that post-editing changes required to improve fluency have less impact on productivity. Since post-editors frequently dismiss MT and post-editing for Fluency issues (word order, word from agreements…), it is highly relevant for our daily work around educating the supply chain.

2.2 Assumptions confirmed

As mentioned, the Adequacy score showed a strong correlation with the productivity delta and gives us an indication of the type of post-editing effort required for the particular program. On the other hand, we found a strong negative correlation between BLEU and PE distance, providing evidence that automatic scores alone cannot be relied upon as a sole indication of the quality of raw MT output.

Among all our language groups, Romance languages render the highest productivity rates. In relation to content, user assistance produces the best productivity rates when publishable quality standards are required. Content types with lower final (i.e. after post-editing) quality expectations like UGC, have even higher productivity gains.

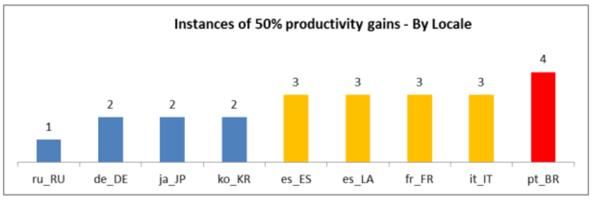

Fig. 2 Instances of productivity gains over 50% by locale, the numbers reflect the quantity of tests that received a score over 50%

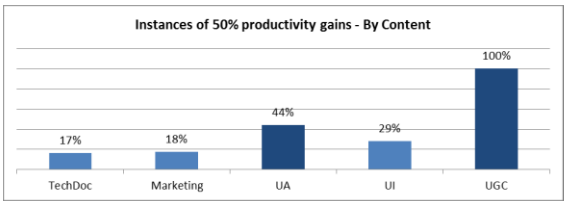

Fig 3 Instances of productivity gains over 50% by content type, the numbers reflect what percentage of tests received a score over 50%

Finally, we could not link negative productivity to a specific content type, even though a traditionally difficult type like marketing was among the content types contained on the dataset.

2.3 Outliers

Throughout the analysis, we observed some results that did not align with our expectations. These findings were particularly interesting to us and we want to focus on them in our presentation, as they give insights into post-editor behavior, variation in input methods, truth and myths regarding best performing languages for MT, etc. The term *outlier* in that sense is here not to be understood as “data to be ignored”, but quite on the contrary, “data to take note of”.[1]

For instance within the group of the above mentioned romance languages, there were still noticeable differences. While Brazilian Portuguese topped the raw MT quality assessments and productivity throughputs (irrespective of the underlying MT system or content type), results for French were a lot less consistent and generally lower.

3 Individual Productivity Influencers

Before talking about variations in individual productivity gains from MT in post-editing, it is important to point out that Adequacy & Fluency scoring exercises, when carried out by several speakers of the same locale on the same content, tend to lead to similar results. Of course, here, too, there is some individual variation, but overall scores tend to move in similar directions, confirming the scores and trends of the other evaluators.

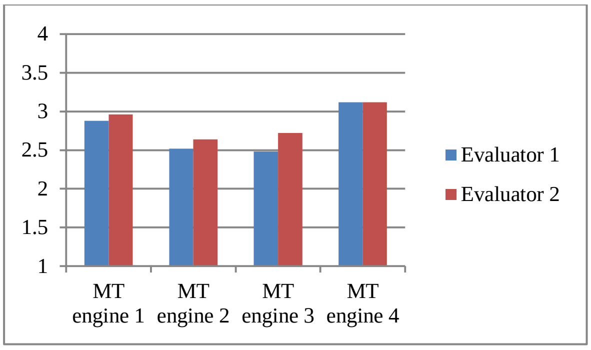

Fig. 4 Accuracy scores of two German evaluators for four different MT engines, using identical sample content. Despite minor variation, trends are the same.

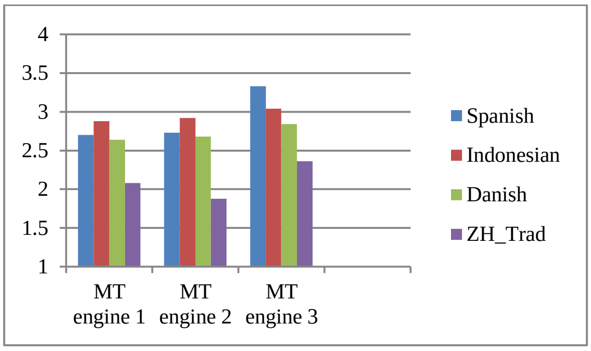

Fig. 5 Human Evaluation scores by evaluators of four different languages for three different systems. The content sample was identical

With productivity gains in productivity tests, however, we see strong variation from one translator to the next. Although some content types do lend themselves better to MT, the correlations were not as clear-cut or within our own expectations (see Marketing earlier on). Language pairs are expected to yield different results with MT, but as the Brazilian / French example shows, are not a sole explanation.

Earlier papers have called out factors such as translators’ experience and technical skills (Guerberof, 2009; Almeida and O’Brien 2010). Verleysen (2013) also mentions translators’ working methods in the European Commission’s Newsletter. While experience and technical skills probably play a part one way or another, they do not as yet show to be consistent factors in our data. Working methods strike us as very interesting and relevant, as the case in 3.1 further suggests.

For some languages (e.g. Romance), trends are more uniform, for others (e.g. German, Russian, Japanese, Korean, as mentioned later on) they vary greatly, making it difficult at times to establish a fair average of what could be the expected productivity gains for this content and language.

With the aim of learning from individual behaviors and predicting future productivity gains, we ask ourselves two questions:

- *What circumstances or variables most reliably facilitate good-quality, highly productive post editing?*
- *Do conditions and parameters outside the post-editor’s control facilitate or hamper his or her success?*

In our analysis of over a hundred cases we noticed that the deviations between individuals are very significant, especially when it comes to MT

[1] We should note here that outliers caused by corrupted data, faulty results, errors in human annotation etc. had been discarded from the database from the outset.

post-editing. It is tempting to assume that the increase between HT and MT is progressive and that every individual improves their performance when they change from translation to post-editing. The reality is not that simple; not all translators benefit from MT output in the same manner and some do not benefit at all.

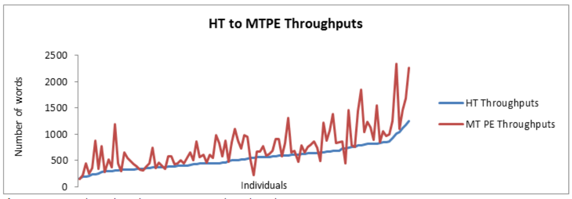

Fig. 6 Translators benefit from MT output in a different way.

In terms of productivity gains, two groups in particular are interesting:

1. Individuals who gain 50% productivity or higher when they move from translation to post-editing.
2. Individuals whose *translation* throughputs are well above the average. We focus on translators who produce 600 or more words per hour.

Our initial analysis has shed some light on potential common characteristics among the first group:

- Language combination: English into Romance languages. Note: Above 50% productivity gains were also seen for Russian, German, Japanese and Korean, but Romance languages (with some internal variation) are still showing higher productivity gains and more consistently so.
- Content type: User Assistance.

But what about the other individuals, the ones who outperform in translation, the ones who can translate at a pace well above the average? Are they able to gain good productivity gains when moving onto the task of post-editing or is there something like a "plateau" in terms of daily individual throughputs? Do they share common characteristics? These are questions we want to further investigate and share first insights at the summit.

Another group of whose translation behavior particularly caught our interest are the English into Japanese translators.

3.1 The Japanese case

Japanese continuously proves to be one of the most challenging locales for MTPE programs, not only with regard to achieving raw MT output of a good quality level.

Through our evaluations and working with a range of translation partners for this locale, we discovered a few aspects how Japanese translators as a group deviate from other languages (e.g. often no formal translation training, very different translation volumes on specific programs compared to FIGS for instance,…) that could potentially influence post-editing productivity. The one that intrigued us most relates to Input Method Editors (IME): it appears that Japanese translators always use some form of IME when working in CAT tools. Some of these IMEs are more elaborate than others, and also some translators are savvier in making best use of them than others. While they certainly have an impact on translation speed, the impact on post-editing speed is not entirely clear to us at this stage, but it is possible that good skills around IME contribute more to productivity for Japanese than MT does.

4 Conclusion

An exhaustive correlation of MT evaluation data was carried out across a wide range of locales, content types and MT systems at our company on 2013. The initial analysis of correlations and data confirmed certain assumptions, but also highlighted the complexity around MT quality and predicting productivity gains, especially with regard to individual translators' behavior.

With regard to translator behavior, there are two areas in particular we would like to analyse further through extensive surveys, in order to share results at the summit: firstly, those translators that already have above average throughputs for translation – how, if, do they benefit from MT? Secondly, IME for Japanese translators: what tools and options are available, what are different levels of sophistication, how are people using them etc., always with a focus on potential impact on post-editing.

Acknowledgement

The authors would like to acknowledge the support of Olga Beregovaya, Andy Mallett, Elaine O'Curran, David Clarke, Karen McLaughlin and Louise Quinn.

References

Guerberof, Ana. 2009. *Productivity and quality in MT post-editing*. MT Summit XII - Workshop: Beyond Translation Memories: New Tools for Translators MT, Ottawa, Ontario, Canada.

De Almeida, Giselle and O'Brien, Sharon. 2010. *Analysing Post-Editing Performance: Correlations with Years of Translation Experience*. Proceedings of the EAMT Summit at St. Raphael.

Verleysen, Piet et al. 2013. MT@Work Conference: by practitioners for practitioners. *European Commission Languages and Translation Newsletter*. Issue #6 on Machine Translation. 6-9.