revista.tradumàtica
tecnologies de la traducció

476-483

Sampling for machine translation evaluation

Rubén de la Fuente
Machine Translation Specialist

PayPal

ABSTRACT

This paper intends to provide an overview of best practices developed within PayPal for designing and preparing samples for different tasks included in the process of machine translation evaluation.

Keywords: Sampling, machine translation, evaluation, QA.

RESUMEN *(Mostreigs per a la l'avaluació de traducció automàtica)*

Aquest article pretén oferir una visió general de les millors pràctiques desenvolupades a PayPal per al disseny i preparació de mostres per a diferents tasques incloses en el procés d'avaluació de la traducció automàtica.

Palabras clave: mostreig, traducció automàtica, avaluació, control de qualitat.

RESUM *(Muestras para la evaluación de traducción automática)*

Este artículo pretende ofrecer una visión general de las mejores prácticas desarrolladas en PayPal para el diseño y preparación de muestras para diferentes tareas incluidas en el proceso de evaluación de la traducción automática.

Paraules clau: muestreo, traducción automática, evaluación, control de calidad.

1. Introduction

In Paypal, we use machine translation as a tool to help our translators keep up with the demanding turnarounds required within an agile development environment while adhering to our quality standards. Before we roll any MT system out to production, we need to test it extensively to make sure it will be useful for our translators. Once in place, we keep monitoring the systems on an on-going basis to make sure they are performing as expected.

One of our challenges then when evaluating machine translation technology is to design and select our samples carefully, so that they are representative of our content, and analyzing them will give us a good understanding of how that particular system is performing and also insights on next steps to take to improve them. We find that for this purpose selecting our samples randomly would be less informative and that's why we resort to systematic sampling, i.e. sampling performed using knowledge of the content, as described in TAUS Sampling Best Practice guidelines (Taus, 2014). In the next sections, we will describe how we acquire that knowledge of the content and how we apply it to sample design. Finally, by way of conclusion, we will share a list of recommendations for sampling preparation based on our experience.

2. Sampling should be adapted to the test

After a few years working with MT, we have found that it is the combination of different tests, and not a single one in particular, that will gives us reliable information about MT performance. The tests we generally perform when evaluating MT systems are the following:

Els continguts de la revista estan subjectes a una llicència Creative Commons (CC BY 3.0)

- Edit distance test, where raw machine translation output is compared automatically against a human reference translation and provides the similarity between the two expressed as a percentage. This test provides insights on the leverage you get from machine translation: the greater the similarity between machine translation output and the reference translation, the less amendments machine translation needs.
- Engine ranking, where human evaluators need to rank the translations produced by typically 2-3 different systems. We perform this test when we are looking for an MT solution for a new language pair or we want to make sure an upgrade to an existing system will bring an improvement in performance.
- Quality evaluation, where human evaluators rate machine translation output in terms of adequacy (meaning is transferred accurately) and fluency (how natural output sounds) and can also categorize errors in output based on predefined categories. Also performed generally when looking for MT solutions for a new language pair or testing upgrades. The error categorization can be particular useful in finding specific areas where an engine can be improved.
- Productivity tests, where human evaluators are to both translate and post-edit comparable segments, and then the difference in productivity is calculated. It can be used to validate the insights from the edit distance tests, i.e. to verify that MT output with a certain degree of similarity with a reference translation allows to increase productivity as compared with translation from scratch.

Different tests have different purposes and requirements, so sampling should be adjusted accordingly. Automated tests, such as edit distance, are less resource-demanding, so it is easier to work with bigger samples. For instance, we generally take all assignments for a given month. On the other hand, manual tests involving human evaluators will take up more resources and hence we'll need to work with smaller samples that will be selected carefully in order to get the insights we need from the test. Generally, we first rely on automated tests as a first exploratory step and only start manual tests if promising results have been achieved on the automated ones.

3. Resources leveraged for sample design

When we design a sample, there are mainly two existing resources that we leverage: monthly metrics and translation memories.

Our monthly metrics automatically collect project name, edit and review distance (edit distance refers to the amount of changes as percentage of the target text performed on MT output by the external translator, while review distance refers to the amount of changes performed by PayPal's in-house linguist on the text submitted by the external translator) and word count for every project that has been translated using MT. In the case of non-MT languages, we also collect word count and review distance, so that we can have a baseline to refer against when MT is enabled. The metrics are collected using an internal tool called MT analyzer, which produces edit and review distance figures (both at segment and project level) as well as a list of the changes undergone in every string. MT Analyzer is scheduled to be run at the end of every week and stores the scores in a SQL database, which is later queried to prepare the monthly metrics. The metrics are used to identify content types with the highest volume and also content types most challenging for machine translation. When preparing the sample, we want to make sure that these content types are duly represented.

Translation memories can be exploited as a corpus to be analyzed in order to identify certain attributes of our content, like sentence length (in words) or the amount of inline tags

Els continguts de la revista estan subjectes a una llicència Creative Commons (CC BY 3.0)

present in segments. The presence of tags is very relevant because they can interfere with the linguistic parsing by the MT engine and hence degrade the MT output.

4. Attributes to be considered

The attributes to be considered when preparing a sample are the following:

- Size, both in terms of number of segments and word count. For automated tests, the more the better, with a minimum of 500 segments. For manual tests, 100-250 strings of carefully selected strings can be enough.

- Content types, which should include both content types with highest word count and edit distance.

Figure 1 shows the word count distribution per content type for European French for the last two quarters, which we calculated from our monthly metrics. Based on this, when we prepare a sample for French, we want to make sure that most of the content is of the product type.

Figure 2 shows the overall edit and review distance for European Spanish for the last two quarters, again calculated from our monthly metrics. Based on this, when we prepare a sample for European Spanish, we want to make sure that particularly challenging content types like mobile are well-represented in the sample.

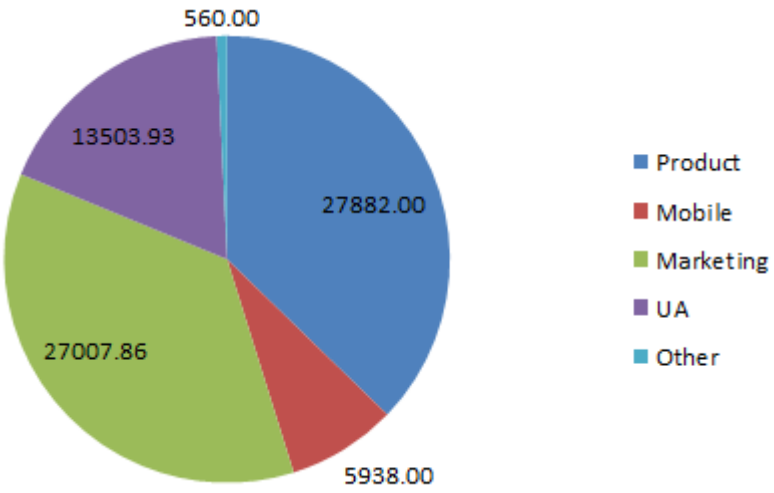

Figure 1: word counts for European French in the period Q4 2013 - Q1 2014

Els continguts de la revista estan subjectes a una llicència Creative Commons (CC BY 3.0)

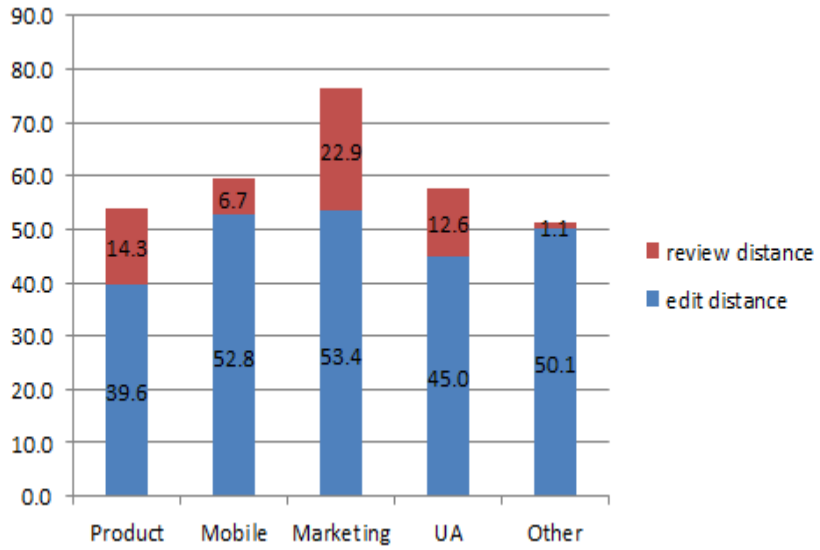

Figure 2: edit and review distance for European Spanish in the period Q4 2013 - Q1 2014

- Segment length distribution. The sample should include examples of most common lengths. Figure 3 shows the sentence length distribution in strings from 2013 in the Brazilian Portuguese TM. Strings were extracted from TM based on timestamp and then word count distribution was calculated in Excel using the formula to count the number of words in a cell or range described in Excel's online help (Office Support). Most strings are in the range 1-20. We take into account the proportion for each bucket when preparing the sample.

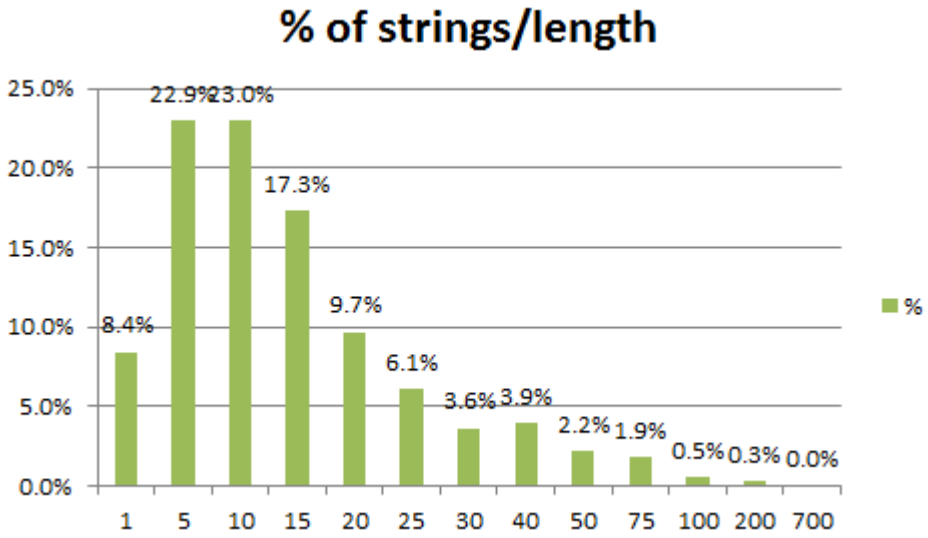

Figure 3: sentence length (words) distribution for Brazilian Portuguese in the period Q4 2013 - Q1 2014

- Tagging distribution. The sample should include examples with most common amount of tags. Figure 4 shows the tag distribution in strings from 2013 in the Brazilian Portuguese TM. Tag distribution was calculated in Excel using the formula to count the occurrences of text, characters, and words described in Microsoft Knowledge Base (Office Support, 2007). Over half of the strings have no tags at all, whereas one third has one or two.

Els continguts de la revista estan subjectes a una llicència Creative Commons (CC BY 3.0)

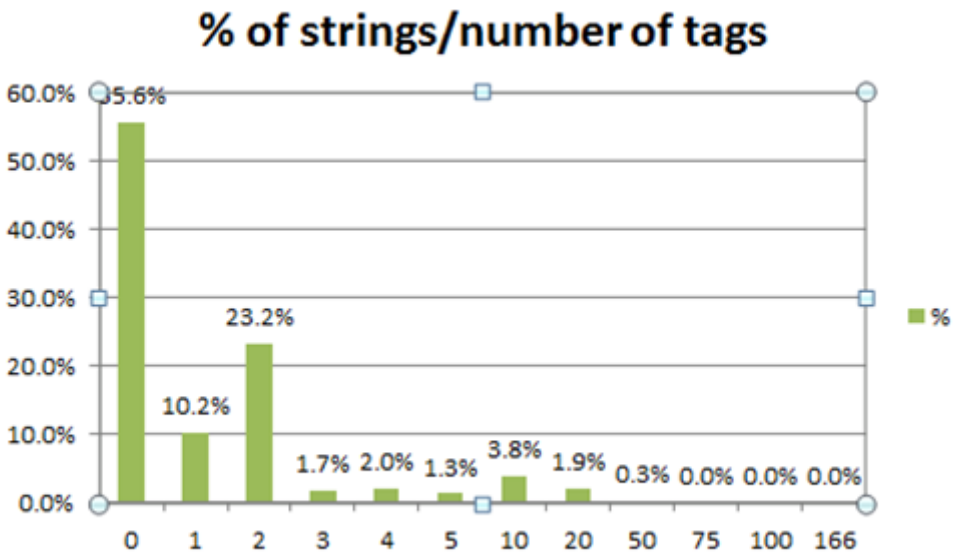

Figure 4: tag density distribution for Brazilian Portuguese in the period Q4 2013 - Q1 2014

5. Sample selection recommendations

Based on the findings presented above, we have come up with a list of recommendations for sample selection:

- Strings should have lengths in 1-20 words, in the same proportion found in the translation memory.

- We should make sure that the sample does not include any duplicate segments, segments where source and target are the same or very similar segments, as they are less informative. The first two checks can be performed in Excel in a semi-automated way, using the EXACT formula (Office Support) and any of Excel's methods for removing duplicates presented on (Kuo, 2013). For the third one, we resort to alphabetical sorting and a manual check.

- If we are ranking different engines, we want to avoid segments that return very similar translations for all engines. Edit distance against a reference translation can be used as a filter. We suggest 5% threshold.

- Maximum of 2 tags per segment allowed.

This set of recommendations is by no means comprehensive (and is hence subject to on-going revision) nor claims to be scientific in nature. It is the result of several years of working with machine translation and intends to serve as a guide in the challenge of being as thorough and efficient as possible in spite of limited resources.

Bibliography

Kuo, M. (2013, October). How to Remove Duplicate Values in Excel. Retrieved November 21, 2014, from mbaexcel: http://www.mbaexcel.com/excel/how-to-remove-duplicate-values-in-excel/

Els continguts de la revista estan subjectes a una llicència Creative Commons (CC BY 3.0)

Office Support. (2007, January 24). Description of formulas to count the occurrences of text, characters, and words in Excel. Retrieved November 21, 2014, from Microsoft: http://support.microsoft.com/kb/213889

Office Support. (n.d.). Count the number of words in a cell or range. Retrieved November 18, 2014, from Microsoft: http://office.microsoft.com/en-ie/excel-help/count-the-number-of-words-in-a-cell-or-range-HA001034625.aspx

Office Support. (n.d.). EXACT. Retrieved November 21, 2014, from Microsoft: http://office.microsoft.com/en-001/excel-help/exact-HP005209081.aspx

Taus. (2014, July). Best practices on Sampling. Retrieved August 21, 2014, from Taus: https://evaluation.taus.net/resources-c/guidelines-c/best-practices-on-sampling

Els continguts de la revista estan subjectes a una llicència Creative Commons (CC BY 3.0)