A detailed investigation of Bias Errors in Post-editing of MT output

Silvio Picinini spicinini@ebay.com
Localization, eBay Inc., San Jose, CA, USA
Nicola Ueffing nueffing@ebay.com
Machine Translation Science Lab, eBay Inc., Kasernenstr. 25, Aachen, Germany

Abstract

The use of post-editing of machine translation output is increasing throughout the language technology community. In this work, we investigate whether the MT system influences the human translator, thereby introducing "bias" and potentially leading to errors in the post-editing. We analyze how often a translator accepts an incorrect suggestion from the MT system and determine different types of bias errors. We carry out quantitative analysis on translations of eCommerce data from English into Portuguese, consisting of 713 segments with about 15k words. We observed a higher-than-expected number of bias errors, about 18 bias errors per 1,000 words. Among the most frequent types of bias error we observed ambiguous modifiers, terminology errors, polysemy, and omissions. The goal of this work is to provide quantitative data about bias errors in post-editing that help indicate the existence of bias. We explore some ideas on how to automate the finding of these error patterns and facilitate the quality assurance of post-editing.

1. Introduction

The use of machine translation (MT) for facilitating the work of translators is increasing throughout the language technology community. The human translator receives an automatically generated translation from the system, and then corrects the errors made by the system. This is called post-editing. As post-editing will gain even more importance, we believe that the quality of this work needs to be evaluated. Translations suggested by MT systems contain errors, and - for several reasons, such as time pressure - the posteditor might leave these MT errors uncorrected. We are calling this effect “bias”, as in the posteditor being “biased” by the MT suggestion, and accepting translation errors.

In our work, we investigated whether the MT system influences the human translator, thereby introducing bias and potentially leading to errors in the post-editing. We analyzed how often a translator accepts an incorrect suggestion from the MT system. Furthermore, we explored the types of bias errors and performed a quantitative analysis.

Our analysis was carried out on translations of eCommerce data from English into Portuguese, consisting of 713 segments with about 15k words. In addition to the MT output and the post-editing, we carefully curated a golden post-editing reference. Using this golden reference, we calculated edit distances and related scores, and then classified and quantified the types of errors that emerged. We observed a higher-than-expected number of bias errors, about 18 bias errors per 1,000 words. Among the most frequent types of bias error we observed ambiguous modifiers, terminology errors, polysemy, and omissions.

The goal of this work is to provide quantitative data about bias errors in post-editing that helps indicate its existence. Additionally, we will provide data about certain types of error patterns that lead to bias. We explore some ideas on how to systematically find these error patterns and facilitate the quality assurance of post-editing. Educating post-editors about bias and about these patterns can help improve the quality of the post-editing work, and therefore the final translation quality delivered to the user.

An early analysis of post-editing of machine translation output is presented in (Krings, 2001). This publication discusses the post-editing process and the quality of machine translations and post-editing, but does not have a quantitative analysis of errors. More recently, (Blain et al., 2011) presents a qualitative analysis of post-editing, focusing on reducing the post-editing effort. In addition to this analysis, the authors present methods for learning corrections from post-editings and improving the MT systems which generated the translations.

2. Analysis of Bias Errors

2.1. Data

We worked on translations from English into Portuguese in the eCommerce domain. The text are descriptions of items which are for sale on the eBay site. The English descriptions, consisting of 713 segments with 15k words in total, were automatically translated using the Microsoft statistical machine translation system, and were post-edited by a human translator, whom we will call post-editor 1 going forward. These post-editings were carefully reviewed by another language expert, whom we will call post-editor 2, who created perfect translations to be used as golden references.

2.2. Methodology

We performed a detailed manual analysis of the post-editings from post-editor 1, comparing them against the golden reference from post-editor 2, in order to detect bias errors. For each error corrected by post-editor 2, we analyzed source, machine translation, and post-editing for potential bias. We classified the errors into certain groups which will be described in section 3.

We used edit distance (Word Error Rate – WER) in two significant ways. First, the distance calculated between the **machine translation and the post-editing**. This is an indication of where post-editing happened and how much. Based on those data, we developed a process (described in a section below) to identify instances of <u>lack of</u> post-editing:

- If the post-editor does not post-edit a segment (for example, by skipping it), the edit distance is zero. This could look like all MT errors were accepted and there was bias, but in reality the posteditor just missed the entire segment. We wanted to find and exclude these instances from the bias analysis.
- If the post-editor rushes through the task and make just one change in a segment, and there were others to make, this will result in a low edit distance. This would look like bias when it is not bias, it is just lack of proper post-editing. We also wanted to find these instances and exclude them.

Second, we used the edit distance between the **golden reference and the post-editing**. The primary use of it was to triage the segments to be analyzed. If the edit distance was zero, this meant that the golden reference agreed in full with the post-editing, so this segment should not be part of the analysis.

The edit distance between post-editing and golden reference can indicate:

- If the edit distance is low, this is an indication that the post-editing was generally good and not many changes were needed.
- If the edit distance is significant:
  - There could be a lack of knowledge – the post-editing made changes and they were wrong, so the golden reference corrected this. This

could appear as high PE-MT distance and also high Golden-PE distance.

- There could be bias – the post-editing accepted the MT and the golden reference changed it. This could appear as lower PE-MT distance and higher Golden-PE distance.

We looked into the numbers for the edit distance through the WER scores, see Table 1. The results are consistent with our expectations: The PE-Golden is higher with bias than without it, which means that there were more corrections for bias segments, as expected. The PE-MT is slightly lower with bias compared to without it, which means that there were fewer changes by post-editors in segments with bias, and therefore they left more errors in them.

avg. WER	All	without bias	with bias
PE vs. golden	0.12	0.09	0.20
PE vs. MT	0.25	0.26	0.21

Table 1. Average WER of post-editing (PE) vs. golden reference and vs. MT output

Finding and excluding content with lack of post-editing

The bias that we are trying to identify happens when the post-editor looks at the machine translation and makes a conscious decision to accept it, and the machine translation is wrong. However, it could happen that the post-editor would skip working on a segment, or could make one change at the beginning of a segment and leave the rest untouched. These would not be example of bias, they would be examples of lack of complete post-editing. In order to try to identify this phenomenon, and exclude it from our analysis of bias, we went through the steps described below.

1. Generated WER scores for each segment, between the post-edited version and the initial MT version (PE-MT).
2. With numbers for each segment, we plotted these numbers on a chart (Figure 1):

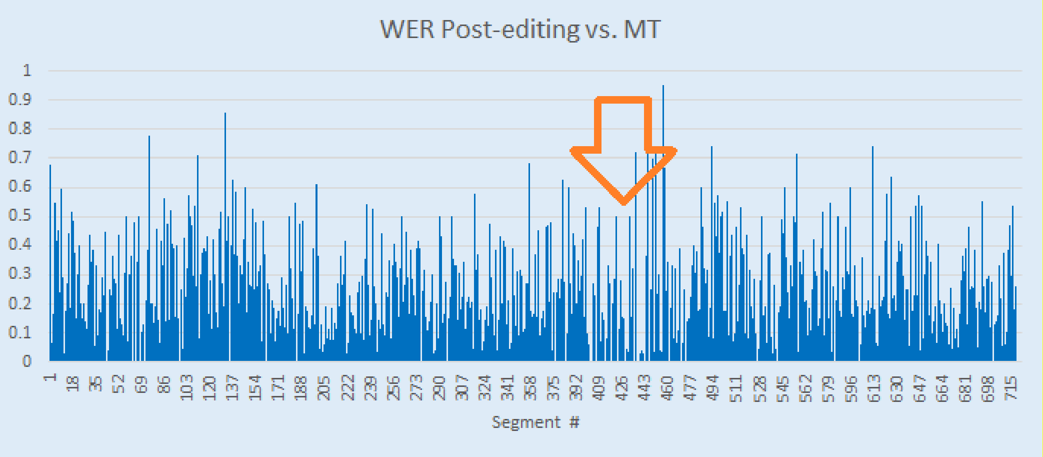

Figure 1. Segment-level WER Post-editing vs. MT

This chart shows regions of data where the volume of post-editing seems lower than the rest of the chart. Further investigation showed that the post-editor indeed failed to do a complete post-editing on segments in this region.

3. We looked for a different chart display that would make this phenomenon more visible than plotting the scores. Therefore, we calculated the average of the WER for the past 30 segments, and plotted this rolling average of distances (shown in Figure 2). The orange line is the average for the file.

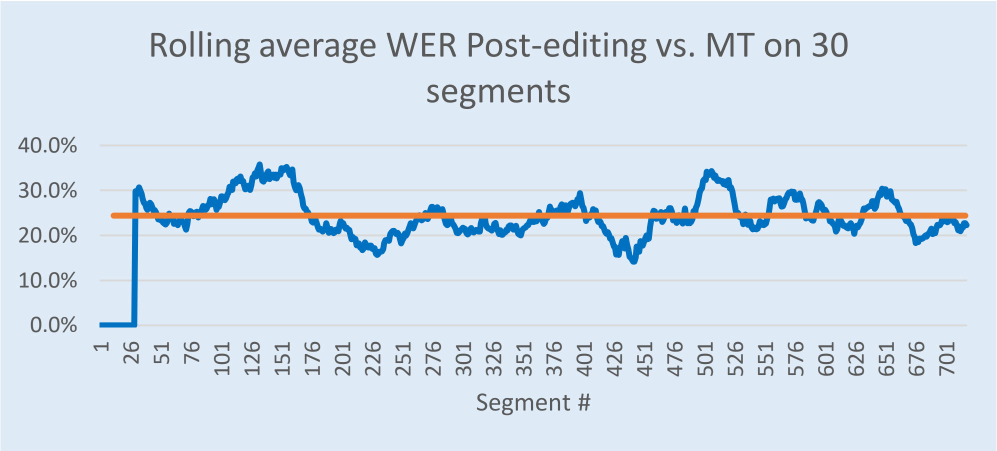

Figure 2. Rolling average WER Post-editing vs. MT on 30 segments

This type of chart shows the amount of post-editing effort progressing through the file. If the post-editor, for example, rushes the work towards the end of the file because of a deadline, this will be reflected in a lower WER/edit distance in a series of segments in that region of the file. This lowering will appear in the chart, as the rolling average will go down for that region. This visualization showed two regions of interest (where the chart shows the lowest values), one region around segment number 221 and the other region around segment number 441. After investigating these regions, we confirmed that the second one had segments lacking post-editing.

4. We looked for one more type of chart and we plotted the “Rolling % of zero-WER in 30 segments” and “Rolling % of low WER (<4%) in 30 segments”. In this chart (shown in Figure 3) we tracked the % of zeros in the past 30 segments. A concentration of segments with no post-editing would start to increase the percentage as we move through them, so regions in this chart with peaks are our regions of interest. We did the same for “% of lows” (shown in Figure 4), tracking not only zero changes but also low % of changes up to 4 %. These are segments that could have changed one character, for example.

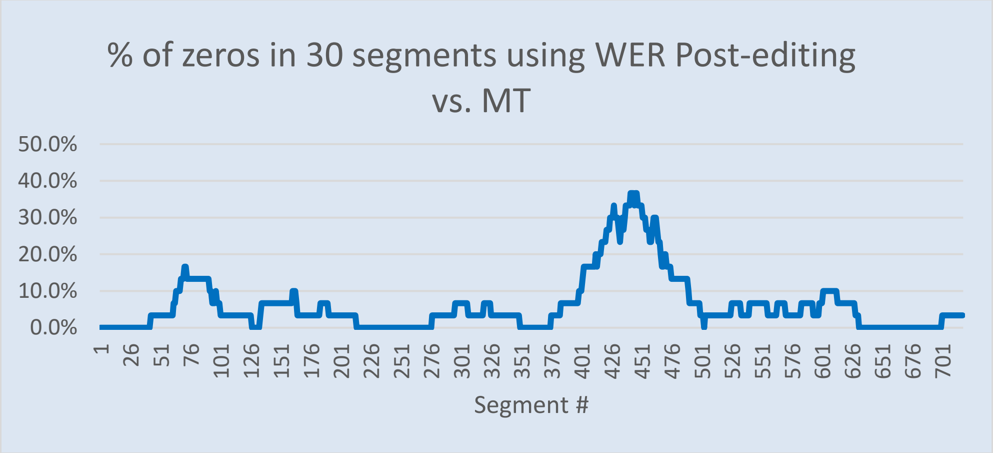

Figure 3. Percentage of zeros in 30 segments using WER Post-editing vs. MT

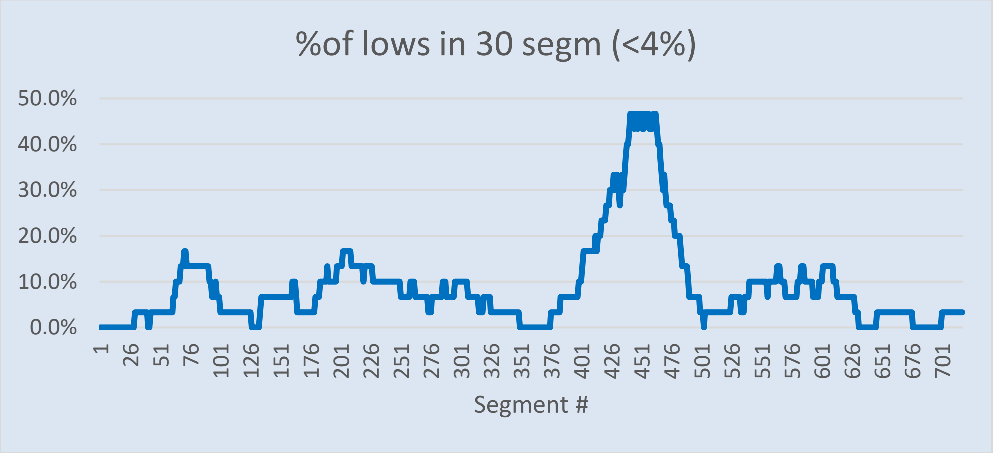

Figure 4. Percentage of low WER (<4%) in 30 segments

Both charts were very effective in pointing out regions of interest (highest values on the chart, around segment # 441 as before. While at first it may seem counter-intuitive to look for the highest numbers when talking about low edit distance, it takes just a few seconds to realize that we are looking for "high concentrations" of low scores, and then the peaks on the charts make sense.

2.3. Findings

General observations

"Is there a significant volume of bias?", that was the question that we wanted to explore for this particular case. While a "Yes" answer can't be easily generalized to other cases, we hope that there is value in concluding that (1) bias happens and (2) this is an issue that needs attention when thinking about improving the post-editing quality.

Types of errors/causes found

Our analysis did not start with a defined set of standard error types. Instead, as error patterns emerged, they became a type. We are used to error typologies, but the classification used in this work is trying to look at **causes of MT errors**. Some of these descriptions below will look more clearly like a cause, such as “Modifiers to Multiple Words" or “Multiword expressions” and others may look like a traditional error type, but there is still a cause behind it. Whatever the causes are, we should just keep in mind that these causes created an MT error, and then bias occurred when that MT error was not changed.

- Multiple Modifiers or Words (MMoW) – this pattern describes situations where a modifier may or may not apply to several words around it. This ambiguity is difficult for the MT to resolve. This more frequently applies to nouns and adjectives, but we opted for a more general name because there are some examples of those, and the same principle applies. Examples of this situation are shown in Figure 5 and 6:

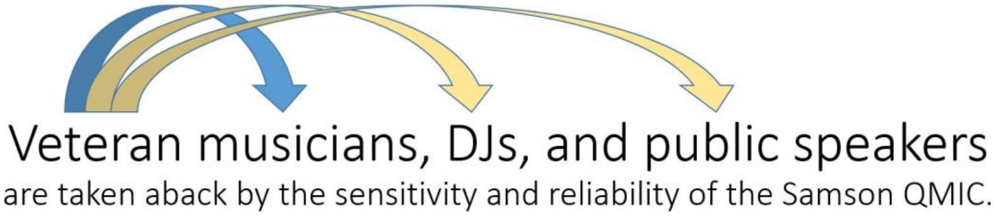
Veteran musicians, DJs, and public speakers
are taken aback by the sensitivity and reliability of the Samson QMIC.

Figure 5. Example 1 of Multiple Modifiers or Words

We as humans intuitively know that this is talking about veteran musicians but also veteran DJs and veteran public speakers. However, the MT engine does not know that, and will produce a translation that says, “DJs, public speakers and veteran musicians are taken…”, and the cause of the error is a modifier adjective that applies to multiple nouns. Another example:

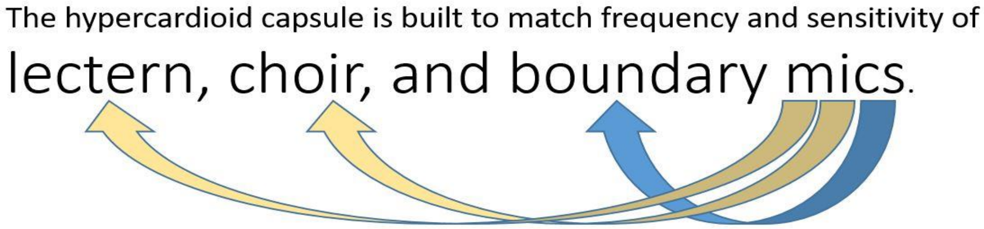
The hypercardioid capsule is built to match frequency and sensitivity of
lectern, choir, and boundary mics.

Figure 6. Example 2 of Multiple Modifiers or Words

In this situation, we have three modifiers applied to one word. We as humans use the context to understand that mics (microphones) probably have frequency and sensitivity and therefore lectern, choir and boundary are three different types of microphones. So this sentence actually means “…lectern mics, choir mics and boundary mics.” The MT does not know that and will produce a translation that sounds like “… sensitivity of boundary mics, and choir and lectern.”, and the cause of the error is multiple modifier (you can see them as nouns or adjectives) that apply to one noun.

This pattern appeared a significant number of times and tends to be difficult for MT. We decided to explore further this pattern in two ways, explained later in this paper:

    - Can we find this pattern more systematically?
    - Does this issue occur also for Neural MT?
- Multiword expressions (MWE) – these are issues where a sequence of words has a completely different meaning than the individual words. Examples of this pattern are idioms and phrasal verbs. It is a difficult construction for the MT to handle because of the change in meaning, so it is a cause of errors made by MT.

Examples include “makes an impression”, “cut short”, “built in”, “turn over to”.

- Polysemy - Polysemous words are words with multiple meanings and therefore multiple translations. In our case, we look at all issues related to polysemous words that have two competing meanings that are popular in the corpora and confuse the MT engine This is a cause of errors for MT.

  Consider, for example, “a choice of restaurants to eat”. The more common meaning of “choice” is probably “to make a choice” but in this example, you are not actually making a choice, and instead the meaning is “variety of options”. If this meaning is less common in the corpora, the MT may make an error. In “performance-conscious photographer”, “conscious” means “photographer concerned with the performance”, but it was translated literally as "did not lose conscience". So the translation ended up sounding like “performance-did-not-pass-out photographer”.

Other examples include:

  - In “Enter a new world of creativity”, “enter” was translated as “insert” as in “enter a password” instead of “walk into” a new world.
  - “fleece-lined compartments” had “lined” translated as “aligned” instead of “covered with fleece”
  - In “Publishes…materials of benefit to the bar”, “bar” refers to lawyers and was translated as the place to go for drinks.
  - In “Washer…including cycles for active wear”, “wear” refers to clothing and was translated with the meaning as in “wear and tear”.
- Mistranslation - In general, “Mistranslation” represents causes that made the MT engine produce a mistranslation. However, every error can be considered a mistranslation. In this work, we classified all possible issues into specific categories. The issues left to be classified as mistranslation are the ones where the translation is wrong but the cause can’t be easily identified. The example of “parent and child” translated as “father and son” should illustrate this category well. We don’t know exactly why the translation is wrong, we only know that it is. This goes into a “Mistranslation” category.

Other Mistranslation examples include:

  - “allow concentration to be focused elsewhere” had “elsewhere” translated literally as a location, when “elsewhere” here means focused on “something else”
  - “reduces eye fatigue and neck pain” had “neck pain” translated as “throat pain”
  - “overcooking” translated as “burned meals”
- Do Not Translate terms – brands and other terms that should not be translated are a cause of errors for MT when the engine has to decide if the term is a brand or a common word. Generic examples would be brands like Gap, Guess or Coach. In our case, examples include the brand Philosophy and a product name called JBL Venue Stadium.
- Terminology – this cause of errors appears when the MT does not know the proper terminology for a certain subject matter. Examples include “focal length” for cameras, “devices” for heraldry, “refrigerator” and “green gas”.

- Part of Speech (POS) –we were interested in this specific cause of error, when the MT would translate a word using the wrong POS for it. Some examples we found showed significant ambiguity that would cause MT errors, some of them difficult even for humans to resolve. Examples include:
  - “Nuts & Bolts component utilities include…” where “component” is an adjective meaning “utilities that compose the Nuts and Bolts…”. The translation treated it as a noun.
  - “The dual apertures of the Vivitar MC Macro Focusing Zoom allow for more flexibility in varying light”, where varying is an adjective meaning “light that can vary”. The translation meant “… more flexibility in the ability to vary light”, treating “varying” as a verb.
  - “One example was gilt -- a process presumably done after striking…”, where gilt is a noun (the action named “gilt”) and it as translated as the adjective “gilt”. In sentences where the structure is “<subject> was xxxx”, the xxxx is usually an adjective, but in our example, it was not.
- Omission of the initial article – it is a common style in English to omit an article at the beginning of a sentence. Examples with and without article include:
  - "AutoCAD LT 2D CAD design software simplifies tasks" vs. “*The* AutoCAD LT 2D CAD design software simplifies tasks"
  - "Zeus IOPS eliminates the wait time" vs. “*The* Zeus IOPS eliminates the wait time"
  - "Familiar six-button configuration provide direct access" vs. “*The / A* Familiar six-button configuration provide direct access"
  - "CenterFlex technology helps enable [...]" vs. “*The* CenterFlex technology helps enable [...]"

While readers of English are used to this construction, the MT notices that pattern and consistently produces translations that miss the initial article in several languages. The impact of that is very different from English because readers of those languages are used to the article being present virtually all the time. This is a systematic cause of MT inadequacy. The bias in post-editing consists of not adding the initial article on the target language.

- Untranslated words – we found instances of words that should have been translated and were not. Examples include: “POI” (acronym for Points of Interest), "non-resonant", "adaptogen", "dot inlay". These issues may be linked to these words being out of vocabulary.
- Omission – we tracked omissions made by MT and not corrected by post-editing. Examples include: card in “card printers”, sharp.
- Addition – same as omission for words added by MT and not removed. Example: added “obtain”.
- Prepositions – significant changes of meaning can be caused by a preposition. A different preposition or its omission in the translation may have an impact. Example:
  - In “save consumer’s money by reducing the operating costs”, the preposition “by” is what defines the meaning as “the reduction of operating costs is what will save consumers money”. The translation sounded as “save consumers money, reducing the operating costs” and actually reversed the meaning as if “save consumers money” would “reduce the operating costs”.

    - "The sole in the Callaway Wedge" was translated as "The sole in wedge shape of the Callaway" changing the meaning just with one preposition.
- Gender agreement – we tracked when the MT made the wrong agreement and was not corrected
- Number agreement – same as gender for singular/plural
- Word order – MT created wrong word orders and they were not corrected.
- Grammar - MT created wrong word orders and they were not corrected.
- Verb tense – MT used the wrong tense for a verb and this was not corrected. Examples include:
    - In "Getting a camera with a greater number of megapixels means cropping and enlarging won't adversely affect picture quality", the gerund "getting" actually does not mean that you are actually already getting the camera at the moment. This would be the meaning for the gerund. Instead, it means the hypothetical action of the infinitive, something like "To get a camera… means cropping… won't affect…". This infinitive is the tense that is required to appear in the translation. The MT will translate as a gerund and the post-editing needs to change to an infinitive. If this does not happen, there is a bias.
    - English has almost no difference between the subjunctive mode and the indicative. The construction "If I were invisible" instead of "If I was invisible" may be the most visible instance of differences in the mode. Yet, there is a popular song that uses the "was" form. This similarity will cause the MT to make errors translating subjunctives as indicatives. If this is not corrected by post-editing, there is a bias. Examples include:
        - "so that they can be lifted" in "the rockets are fitted with magnets so that they can be lifted and loaded with cranes"
        - "(would) freeze herself" and "would cause" in "It seems as if Hilda, while trying to scare up a dancing partner, accidentally freezes herself and causes objects to fly throughout the house"
        - "to be found" in "Usually the puzzles require items to be found and then executed". It means "the puzzle requires that items be found", and not "requires items that will be found".
- Spelling (including language rules) – Brazilian Portuguese had a spelling reform. Therefore, there are new language rules in place. The corpora used to train the MT engine contains content created before the reform. Therefore, there are spelling errors training the MT, and they will appear in the translation. If they are not corrected, there will be bias.

Spelling reforms and corpora for MT will pose a certain challenge for MT systems. Many languages use corpora from different flavors of the language, such as European Portuguese and Spanish versus those used for Brazil and Latin America. The spelling of Portuguese varies a little bit between Brazil and Portugal, so the MT ends up making a few Portugal suggestions for Brazil and Spain suggestions for Mexico or Colombia. The advantages of having more corpora by doing this outweigh the downsides of it. If the corpora will not be fixed, the role of post-editing will include correcting these "cross-border" spelling issues.

Errors per 1k words

The main number that we obtained was the number of bias errors per 1k words. Although there is no official standard for the industry, we typically consider translations as high quality if the number of errors per 1,000 words does not exceed 2, based on our own personal experience. The total number that we found in this work is given in Table 2.

Total Number of words	14,986
Total Number of bias errors	270
Errors per 1k words	18.02

Table 2. Number of bias error vs. number of words in post-editing

This number is about nine times a reference for quality, indicating that the total number of bias errors is significant. We should keep in mind that these are only bias errors. In addition to these, there are regular non-bias errors, where the post-editor makes changes and they are still not correct. The entire picture of quality is comprised of bias + non-bias errors.

Numbers for each type of error

	Polysemy	Mistranslation	Multiple Modifiers or Words	Multiword Expressions	Omission initial article	Do Not Translate terms	Untranslated	Omission	Addition	Terminology	Gender agreem	Number agreem	Prepositions	Word order	Spelling (incl Lang Rules)	Grammar Verb tense	Part-of-Speech	Total
Number of errors	54	56	22	14	10	9	15	16	4	14	7	12	8	8	4	9	8	270
Errors per 1k words	3.60	3.74	1.47	0.93	0.67	0.60	1.00	1.07	0.27	0.93	0.47	0.80	0.53	0.53	0.27	0.60	0.53	18.02

Table 3. Numbers for each type of error

Table 3 lists the frequency of each of the error types which we defined during our analysis. The breakdown per type shows that several types of causes of errors (described previously) are deserving of further attention:

- Polysemous words
- Multiple modifiers or words
- Multiword expressions
- Terminology
- Omissions
- Untranslated words

- Other causes of mistranslation

Detailed Analysis for Errors caused by Multiple Modifiers or Words (MMoW)

We analyzed the error caused by Multiple Modifiers or Words in more detail. We found 22 instances of this error, indicating about 1.5 errors per 1k words. This is a significant number for just one type of error.

Next, we were interested in the question whether this error occurred with the same frequency for different types of MT systems. Therefore, we compared the output from 2 different types of MT systems for these 22 segments to find out whether these errors stem from inherent complexity of the source segment. We used this issue to have a sense of the impact that Neural Machine Translation may have on the quality. The hypothesis is that if NMT produces an output that contains less causes of errors, there will be less errors and therefore less bias. We wanted to see if this happened.

We looked into the 22 issues identified originally on Microsoft Statistical MT and created a NMT output from Microsoft Neural MT for them. We then evaluated to see how many of the original 22 errors were still present in the NMT output. We found that 10 out of 22 times, the NMT system corrected the error, meaning that it improved over the SMT system in 45% of the cases. However, in the remaining 10 segments, we still observed the same type of error in the NMT output. These results indicate that NMT tends to produce less errors than SMT for the post-editing corrections. This leads to better post-editing quality. However, 55% of the errors in the SMT output were still present in the NMT, indicating that the issues that a significant portion of the issues that are difficult for SMT remain an issue for NMT. This seems to indicate that the work on patterns that we started here would be a worth pursuit in improving NMT, and in evaluating it.

Once we identify that Multiple Modifiers or Words was an issue worth our attention, we thought of how we could find these expressions in a more semi-automated way. The process that we used can be described in these general steps:

1. Run a POS tagging of the source content

2. List tokens and tags and simplify the POS tags to a minimum; see Table 4 for examples. Create patterns indicating errors and find these patterns in the tagged content

We applied a simple formula to find a pattern: adjective-noun-noun and found the example “enhanced telephony capability” above. We also looked for another pattern: adjective or noun-noun-“and”-noun. We found examples such as “cook time and temperature” with this pattern. Analyzing the data, we found that:

- Using only these two narrow formulas we already found 7 out of 22 issues (32%). This indicates that a few formulas could find the majority of the patterns.

Token	**Tag**	**Simplified Tag**
the	DT,B-NP-plural	DT
enhanced	JJ,enhance/VBD,enhance/VBN,I-NP-plural	JJ
telephony	NN:U,I-NP-plural	NN
capability	NNS,E-NP-plural	NN

Table 4. Examples of POS tags and simplified tags for English tokens

- We manually analyzed the 22 MMoW issues to find out how many were suitable to be found with formulas/patterns. Out of 22, 20 of them could be found. This indicates that most of the MMoW issues are findable with patterns, and that there is potential to semi-automate the harvesting of these terms from a content tagged with POS.

3. Conclusions

1. We found significant bias in the post-editing of MT. This cannot be generalized to all cases, but it shows that the bias exists and is an issue to be considered as part of improving post-editing.
2. We found patterns that cause MT errors and can cause significant bias. These patterns should be considered for improvement of post-editing and for measurement of post-editing quality.
3. We found that it is possible to apply some automation in detecting the error patterns that cause errors on MT.
4. We found that Neural MT is likely to reduce the errors from bias by eliminating the original MT error. However, a significant percentage of the issues that cause errors on MT are not resolved by Neural MT and remain of interest for improving and measuring the quality of Neural MT.

4. Future Work

1. The semi-automated finding of patterns should be explored further. Once a representative number of instances of patterns is obtained, different metrics can be calculated. For example, we could find that there are 100 instances of MMoW in the content. If, upon reviewing them, we find, for example, 43 errors, this indicates that this type of error is produced by MT 43% of the time.
2. We think that there is potential in measuring the quality of the MT output based on difficult issues instead of a random sample. If a system 1 performs better than another system 2 on polysemous words, multiple modifiers or words, and multiword expression, it is likely that this system 1 will perform better on any translation than system 2. The same reasoning of measuring difficult words can be applied to measuring post-editing quality. We would like to create a measurement method that is not based on random sampling nor error typology, that targets difficult words, that is not subjective (make simple binary decisions), that is fast, cost-effective and suitable for crowdsourcing (with bilingual people and not professional linguists). We are working on this topic.

References

Blain, Frédéric, Senellart, Jean, Schwenk, Holger, Plitt, Mirko and Roturier, Johannes (2011). Qualitative analysis of post-editing for high quality machine translation. In *Proceedings of MT Summit XIII: the Thirteenth Machine Translation Summit*, pages 164-171, Xiamen, China.

Krings, Hans P. (2001). *Repairing texts: empirical investigations of machine translation post-editing processes*. Vol. 5. Kent State University Press.