What's Your Pick: RbMT, SMT or Hybrid?

Catherine Dove
PayPal
cdove@paypal.com

Olga Loskutova
PayPal
oloskutova@paypal.com

Ruben de la Fuente
PayPal
rdelafuente@paypal.com

Abstract

All types of Machine Translation technologies have pros and cons. At PayPal, we have been working with MT for 3 years (2 of them in a production environment). The aim of this paper is to share our experience and discuss strengths and weaknesses for Rule-based Machine Translation, Statistical Machine Translation and Hybrid Machine Translation. We will also share pointers for successful implementation of any of these technologies.

1 Introduction

PayPal works with a 3-week release cycle and does simultaneous shipping for all languages. Our motivation to use machine translation is to achieve timely deliveries with the highest possible level of quality (cost savings are also considered but have not been the main goal). The MT engines are connected to our translation management system and every segment under 75% match is machine-translated. We do full post-editing of all the MT output and are not considering to use as-is for any purpose as of now.

The source content for machine translation is mainly dynamic html, using variables that are replaced by corresponding value at run-time. The challenge of these tags for machine translation is two-fold: 1) tags can interfere with the parsing of the sentence, resulting in faulty output, and 2) they might require adjustment of surrounding text or reordering depending on their value.

We currently have MT engines in place for French, Italian, German, Spanish, Russian, Danish, Norwegian, Swedish and Simplified Chinese. We also have a normalization engine converting from US English to British and Australian English.

2 Indicators for evaluating and selecting Machine Translation engines

We have identified two main indicators for evaluating and selecting MT engines: linguistic quality and ease of integration with the existing CAT tools and workflows. Often, all the focus is set on the first indicator, but I'd say they are both equally important: an engine producing great linguistic quality but with a cumbersome integration with existing tools will be of little use.

As for linguistic quality, it is important to be systematic and go beyond showing a sample to an in-house linguist or an external vendor and ask for feedback. Please bear in mind that linguistic quality does not mean the MT output will be flawless, but rather that it will take you less time to post-edit the MT output than to translate from scratch (tentatively, post-editing throughput should be 5000-10000 words/day). To test if this requirement is met, the best is to ask a linguist to post-edit a representative sample (around 1000 words), track the time spent and calculate the edit distance. Symeval (Lavie, 2010) and PET (Aziz, 2012) can provide good insights about post-editing productivity.

In terms of ease of integration, the engine must support seamless communication with your translation management system via API or similar.

3 Rule-based Machine Translation

We first started with RbMT for Russian, Spanish, French, Italian and German. Our vendor took care of the initial customization of rules and user dictionaries and then this task was allocated to in-

house linguists. In-house linguists update the user dictionary with new terminology based on forecast projects created with draft files. By updating the user dictionary before the actual production job takes place, we expect to significantly reduce the post-editing effort.

In order to handle the tags, we have included them in the user dictionary with the appropriate part of speech. The RbMT engine takes care of adjusting the surrounding text and reordering the tags if needed.

To ensure good performance of the engine, we keep track of the edit distance (i.e. the amount of rework needed to bring the raw MT output to publishing quality) and the review distance (the rework of translation from vendor by our in-house linguists). The figures are presented in the table below. The metric used is WER (Och, 2003), since we find it more intuitive than BLEU.

Language	Edit distance	Review distance	Volume (words)
French	46.33%	9.1%	38900
Italian	49.05%	16.94%	40149
Spanish	33.67%	6.30%	56269
Simplified Chinese	54.43%	2.69%	80367

Table 1: edit/review distance for RbMT languages

The main advantages of RbMT in our experience are the ease of customization (linguists can start playing around with the tool after a few hours of training; user dictionaries can be amended on the fly to fix errors), the good handling of tags and the predictability (with a basic understanding of the tool, you know what kind of output you can expect).

The main disadvantages are that they require more manual work (3 hours/week per language on average) and that the output will be accurate and grammatically correct, but sometimes not very fluent. They are also more expensive than statistical MT engines.

4 Statistical Machine Translation

We started looking into statistical machine translation because RbMT was not available with the vendor for new languages we needed, namely Danish, Norwegian and Swedish.

An external vendor took care of building a SMT engine for us, using our translation memories as training corpora. The technology is based on open-source toolkit Moses (Koehn et al. 2007). The idea of using Moses out-of-the-box and avoid vendor costs can be tempting. However, Moses is rather complex and vendors have customized it further for better results with language pairs other than the ones the system initially supported.

The edit and review distance figures are presented in the table below.

Language	Edit distance	Review distance	Volume (words)
Danish	36.18%	0.32%	89747
Norwegian	37.19%	N/A	105674
Swedish	43.56	2.41%	115544

Table 2: edit/review distance for SMT languages

What makes SMT engines really appealing is that they require no customization work on the part of linguists: the engine learns by itself through statistical analysis of translation memory corpora. Therefore, a lot of effort in terms of manual work and training is saved, and the engine can be ready in a matter of days. They are also generally cheaper than RbMT engines. A word of caution about SMT: the quality of the output is going to be only as good as the quality of the translations in your TM. If you are concerned that your translation memory might have terminology inconsistencies or mistranslations, it is best to do some QA on it before starting the engine training.

The drawback of SMT is that for the time being they are only capable of keeping the tags in place (without reordering or adjusting of surrounding text). This is not a limitation of SMT itself, but rather of the translation memory corpora, that replace tag content (e.g. <result>countryName</result>) with numeric placeholders ({1}), preventing the engine from learning how the tags need to be adapted in the translation. We are currently considering the feasibility of using xliff files (which do contain tag content) instead of tmx files as training corpora. Another limitation is that it is only efficient to re-train them every 3-6 months, so they are not as flexible to incorporate quick terminology fixes as RbMT. In terms of linguistic quality, the most frequent issues relate to wrong word forms, capitalization and punctuation.

5 Hybrid Machine Translation

Our hybrid machine translation uses RbMT output as baseline, then refines it through comparison against a language model created with SMT techniques. Same as with SMT, it is key that the translation memories used as training corpora are in good shape for succesful implementation.

We first started hybrid for German and Russian. Edit and review distance figures are listed below. The edit distance for German and Russian are a bit high due to the rich morphology of these languages (case inflection) and some heavy locale-specific customization requirements. A roll-out to hybrid for Spanish, French and Italian has taken place recently, but no figures are available as of yet.

Language	Edit distance	Review distance	Volume (words)
German	77.88%	13.77%	17269
Russian	69.11%	5.85%	33764

Table 3: edit/review distance for hybrid languages

Initial testing of the hybrid output shows that it is generally more fluent and natural than RbMT. However, some typical mistakes of hybrid include: deprecated terminology (due to outdated translations present in the translation memory), part-of-speech agreement mistakes, extra words in translation, extra punctuation and wrong capitalization (resulting from unpredictable behavior of the statistical component; we've seen these issues reduced in German and Russian after re-training the engine on cleaner data). In spite of these drawbacks, edit distance should be lower overall. Unfortunately, we don't have figures to prove that point at the time of writing this paper.

6 Conclusions

The type of technology that best suits your needs will change depending on the language pair. The organization and resources of your company (both in terms of headcount and existing linguist assets) is also an important factor to consider.

Statistical MT will deliver good results for language pairs in which the target does not have very rich morphology features (Danish, Norwegian, Swedish). For more complex languages (Russian, German) it is worthwhile to invest in hybrid systems. RbMT can deliver good results provided customization is done on a regular basis, but it seems to be less efficient than the other two types of technology.

A good MT strategy should be technology-agnostic and look for the most efficient solution on a case-by-case basis.

References

Aziz, W.; Sousa, S. C. M.; Specia, L. (2012). PET: a tool for post-editing and assessing machine translation. In The Eighth International Conference on Language Resources and Evaluation, LREC '12, Istanbul, Turkey. May 2012.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico,M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). *Moses: Open source toolkit for statistical machine translation.* In Proceedings of the Annual Meeting of the Association for Computational Linguistics, demonstation session.

Lavie, A. *Evaluating the output of machine translation systems.* AMTA Tutorial, 2010.

Och, Franz Josef. *Minimum Error Rate Training for Statistical Machine Translation.* ACL, 2003.