Calculating the Percentage Reduction in Translator Effort when using Machine Translation

Andrzej Zydroń
XTM International Ltd.,
UK
azydron@xtm-intl.com

Qun Liu
Dublin City University,
Ireland
qliu@computing.dcu.ie

1 Introduction

At present there is no precise indication of the benefits of using Statistical Machine Translation (SMT) for potential users. The question ‘is this going to save me time and/or money’ and if so how much, is not addressed in any systematic way. The common answer provided by most SMT service providers is ‘well, it depends’. This is far from the answer that users need to make an informed decision about whether to go ahead with SMT.

What is lacking in the industry today is a description of the main factors affecting the quality of SMT output and how you can use them to provide an indication of the savings that SMT will provide. In the end, the decision on whether to use SMT depends on the amount of time saved during translation. This paper provides a clear indication of the savings you can expect, depending on the key factors that affect the quality of the SMT, based on a simple calculation that provides a Percentage Reduction in Translator Effort (PRTE) that can be expected for a given localization project.

2 Translation Cost

Translation forms part of the cost of localization, and it is often all too easy to forget about the other elements of the overall localization process and subsequent costs. In fact translation itself typically accounts for only between 30% to 50% of the overall cost of a localization project, depending on how much automation is involved in the overall localization workflow. The following diagram shows the standard cost model for a manual localization process:

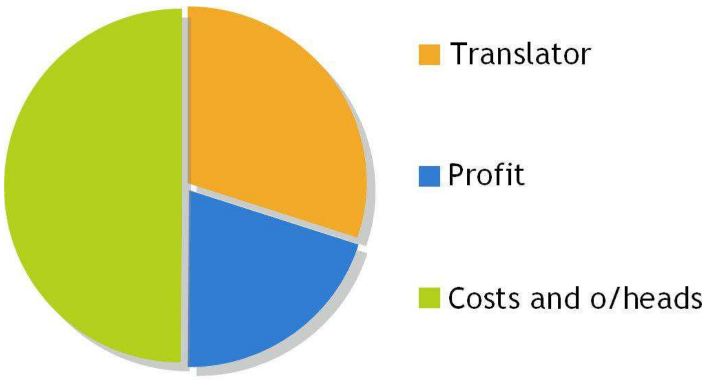

Figure 1: typical localization cost breakdown Prof. Reinhard Schäler ASLIB 2002

As can be seen from the diagram translation itself forms only part of the cost of localization. The other costs, apart from the profit made by the localization service provider, are the management and administrative costs, as well as proofreading, review and correction. An automated translation management system (TMS) can significantly reduce the administrative and management costs of the localization process.

Proceedings of the 38th Conference Translating and the Computer, pages 164–169,
London, UK, November 17-18, 2016. ©2016 AsLing

3 PRTE Calculation

Having put the cost of translation into perspective we can now look at the factors that affect the quality of SMT and consequently the Percentage Reduction in Translator Effort (PRTE).

PRTE can be defined as: The percentage reduction in translator effort by using SMT compared to human translation on its own.

PRTE is the key factor that decides how much savings you can expect to gain from SMT for a given project. The quality of SMT is governed by three major factors:

1. The language closeness (LC): the similarity of the source and target languages in terms of morphology, word order and grammar
2. The amount of training data
3. The relevance of the training data to the current text being translated

If we provide mathematical weightings to these factors we can use them very effectively to provide a calculation of the percentage translator productivity we can expect to achieve using SMT. In order to provide a percentage, we will use a probability type estimation for each factor with a range of 0 to 1, where the value '1' assumes an idealized perfect situation and '0' the opposite.

Let us now consider these factors in detail:

Language Closeness

SMT output is affected by the by the differences between the source and target languages in terms of various aspects, including grammar, word order and morphologies. To put it simply, the closer the two languages are in terms of grammar and word order and morphology, then the better the outcome. To take an extreme case, of say, US English to UK English we can state that the LC is '1.0' as the two variations of English only differ in some spelling instances. Using English as the source again and this time French as the target we can assume a LC value of 0.8, as both languages have similar primitive morphologies and word order. For English to German, we would use a value of 0.6 as the differences in morphology and word order are much more pronounced. For English to Russian or Polish the proposed value would be 0.45 and for English to Japanese it would be 0.25, as there are significant differences in word order and morphology between the two languages.

A good indication of the difference in language models can be found at: http://esl.fis.edu/grammar/langdiff/ - this site provides a comparison for some major languages concerning the difficulties that native speakers of those languages have in learning English. The degree to which these students have issues with learning English is also indicative of the basic differences in grammar and morphology between their native tongue and English and also indicative of the difficulties posed in terms of SMT between English and those languages.

The following table provides an indication of the types of factor where English is the source language. The factors have been arrived at from personal experience and should require further investigation, but they are a good starting point:

Language Closeness factors relative to English	
French	0.800
Spanish	0.775
Portuguese	0.775
Italian	0.760
Dutch	0.750
Swedish	0.700
Danish	0.650
German	0.600
Arabic	0.600
Korean	0.500
Finnish	0.500
Hungarian	0.500
Turkish	0.500
Polish	0.450
Russian	0.450
Czech	0.450
Slovak	0.450
Chinese	0.400
Japanese	0.250

Table 1. LC factors relative to English

If all other factors affecting SMT quality are in an ideal state, then the expected productivity improvement, where the LC is the only factor, then the following graph shows the expected productivity improvement where English is the source language, depending on the target language:

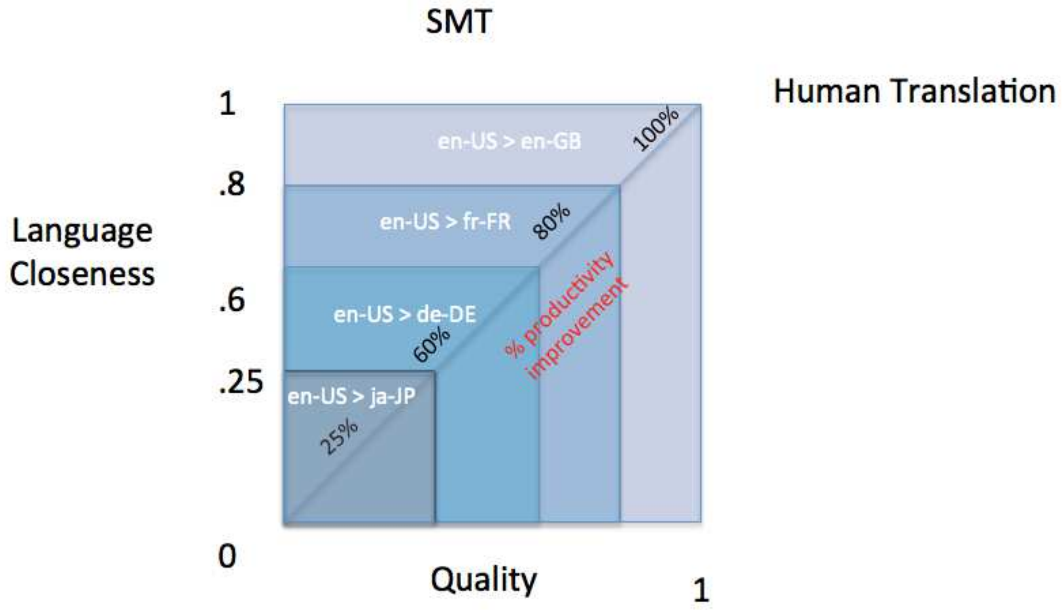

Figure 2: Idealized PRTE for SMT only considering LC factors

Training Set Size Factor (TSSF)

The next key factor regarding SMT quality is the size of the training set. Too small a TSS and there will not be enough data to provide an adequate model for translation. When there is no training data, the TSS should be 0 As the size of data increases the TSS approaches 1, so when the TSS is 1 there is infinite training data. We use the equation below to estimate TSS:

$$TSSF = 1 - 2^{-\frac{Size}{Size'}}$$

Where $Size$ is the actual training data size and $Size'$ is an empirical value which makes TSSF equal 0.5.

What this means is that a training set size of $Size'$ would result in a reduction of the translation effort of 50%. In practical terms this would normally equate to around 50,000 segments, depending on the material being translated. A training set size of 10,000 segments would produce a TSSF of .067 whereas 100,000 segments would result in a TSSF of .75 and 200,000 segments would produce a TSSF of .9375.

The training set size parameters can be adjusted according to the specific requirements of the scenario and how much training data is actually available as opposed to the theoretical optimal amount.

Using the above assumptions, as a very rough rule of thumb normally, you can assume that an optimal training set size of 250,000+ segments would provide a TSS value of approaching 1. Anything less would result in reducing the TSS value roughly by 0.1 for every reduction of 25,000 segments in the training set size.

A constant problem with SMT is the issue of out of vocabulary (OOV) words: these are words that have not been encountered previously in the training set. If the training set size is too small then you can expect a commensurate increase in OOV word instances and therefore more work for the translator.

For the purposes of the PRTE calculation we can assume again a value of between 1 (ideal training set size) and 0 (no training set). Zero would be improbable value (we would not be able to build a SMT engine with no training data), but we can see that if not enough training data is available it would have significant impact on the quality of the SMT.

Domain Similarity (DMS)

Empirical evidence has shown that the quality if SMT also depends on the quality of the training set. A smaller training set on the same topic domain produces much better results than using a generalized training set. Specific domains have their own vocabulary and phraseology that cannot be rendered with a general SMT engine.

For the purposes of the PRTE calculation we can assume a value between 1 (exactly the same specific domain from data for exactly the same organization) and 0 a completely unrelated specific domain. A generic SMT engine would rate 0.25 where the subject matter being translated related to a highly specific domain with its own detailed terminology.

PRTE Formula

The PRTE formula itself takes all three of the above factors to provide an overall calculation that is easy to implement:

$$PRTE = (LC \times TSSF \times DMS) \times 100\%$$

Figure 3: PRTE formula

This can be represented by a three dimensional graph as follows:

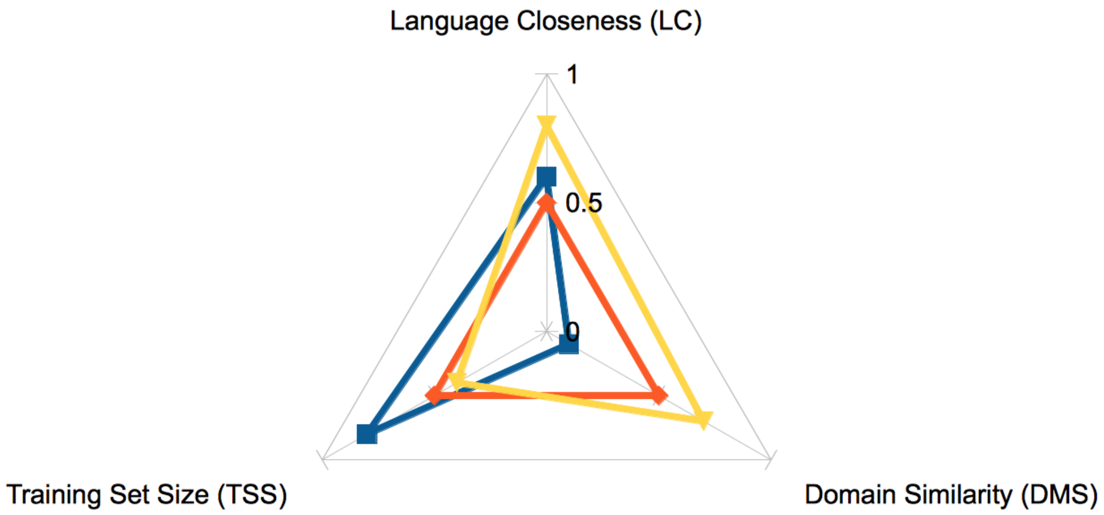

Figure 4: PRTE 3 dimensional graph for various PRTE calculations

To test the validity of the formula we can try some examples:

1. Translating from en-US to en-GB we can assume a LC[1] value of 1. If we have an ideal reference TSSF[2] of 1 and an ideal DMS[3] of 1, we arrive at a PRTE of:

   1x1x1x100 = 100%

[1] Language Closeness
[2] Training Set Size Factor
[3] Domain Similarity

This would mean that the SMT[4] output should require no translator intervention providing a productivity figure of 100%.

2. Translating from en-US to fr-FR we can assume a LC[1] value of 0.8. If we have a slightly less that ideal TSSF[2] of 0.75 but with an ideal DMS[3] of 1, we arrive at a PRTE of:

   0.8x0.75x1x100 = 60%

   This would mean that we should expect an improvement regarding translator productivity of 60% compared with a completely manual human translation.
3. Translating from en-US to ja-JP we can assume a LC[1] value of 0.2. If we have an ideal TSSF[2] value of 1 and an ideal DMS[3] of 1, we arrive at a PRTE value of:

   0.2x1x1x100 = 20%

This would provide an estimated 20% improvement in translator productivity.

4 Conclusion

The PRTE formula is not designed to be a hard and fast assessment of the expected percentage reduction in translator effort, but rather an overall rough estimation of what can be expected. Some of the figures are expected to be at best a ‘guess’ as regards the DMS and TSS figures. The LC values are also a rough approximation and some SMT systems with an appropriate amount of tuning will be able to provide better values. It also does not take into account the differences between individual SMT engines: some will inevitably be better than others. The amount of manual tuning also needs to be taken into account as it requires the input of highly skilled engineers.

Nevertheless the PRTE formula provides a guide to what is achievable for a given situation and roughly an idea of the returns that can be expected. This is vastly better than nothing, or ‘well it depends’ which is the current situation.

[4] Statistical Machine Translation