A Case Study of German into English by Machine Translation: to Evaluate Moses using Moses for Mere Mortals

Roger Haycock
Haycock Technical Services
Purley Rise LE12 9JT
rhaycock@theiet.org

Abstract

This paper evaluates the usefulness of Moses, an open source statistical machine translation (SMT) engine, for professional translators and post editors. It takes a look behind the scenes at the workings of Moses and reports on experiments to investigate how translators can contribute to advances in the use of SMT as a tool. In particular the difference in quality of output was compared as the amount of training data was increased using four SMT engines.

This small study works with the German-English language pair to investigate the difficulty of building a personal SMT engine on a PC with no connection to the Internet to overcome the problems of confidentiality and security that prevent the use of online tools. The paper reports on the ease of installing Moses on an Ubuntu PC using Moses for Mere Mortals. Translations were compared using the Bleu metric and human evaluation.

Introduction

Pym (2012) considers that translators are destined to become post-editors because the amalgamation of statistical machine translation (SMT) into translation memory (TM) suites will cause changes to the skills required by translators. He believes that machine translation (MT) systems are improving with use and a virtuous circle should result. However, free online MT, for example Google Translate (GT), could lead to a vicious circle caused by the recycling of poor unedited translations. Technologists have a blind faith in the quality of translations used as 'gold standards' and Bowker (2005) found that TM users tend to accept matches without any critical checks. Further, the re-use of short sentences leads to inconsistent terminology, lexical anaphora, deictic errors and instances where the meaning is not as foreseen in the original. Pym (2010, p.127) suggests that this could be avoided if each organisation has its own in-house SMT system.

There is another compelling reason for in-house SMT. Achim Klabunde (2014), Data Protection Supervisor at the EU warns against using free translation services on the Internet. He asserts that someone is paying for what appears to be a free service. It is likely that users pay by providing their personal data. Translators using these services may however be in breach of confidentiality agreements because the data could be harvested by others and recycled in their translations. Googles terms and conditions are clear that they could use any content submitted to their services (Google, 2014).

The increased volume of translation caused by localisation does, however, call for automation (Carson-Berndsen et al, 2010, p.53). Advances in computer power have enhanced MT and good quality full post editing can be included in TM for segments (sentences, paragraphs or sentence-like units eg. headings, titles or elements in a list) where no match or fuzzy match is available (Garcia, 2011, p.218). TM developers now offer the facility to generate MT matches to freelance translators, eg. Google Translator Toolkit, Smartcat, and Lilt. This technology will increase the rate of production and according to Garcia (2011, p.228) industry expects that post-editing, with experienced post-editors and in-domain trained engines, for publication should be able to process 5000 words a day. Pym (2012, p.2) maintains that post-editors will require excellent target language (TL) skills, good subject

Proceedings of the 38th Conference Translating and the Computer, pages 100–112,
London, UK, November 17-18, 2016. ©2016 AsLing

knowledge but only weak source language (SL) skills. For this reason I elected to work from German, my weaker foreign language, into English my native language. I have used Lilt to post-edit translated segments to help with evaluation of the SMT as will be explained later.

This paper reports a case study that used Moses for Mere Mortals (MMM) to investigate how difficult it might be for a freelance translator to incorporate an in-house SMT engine into a single workstation by building four distinct Moses engines with different amounts of data. Following an overview of the project the method followed to install, create, train and use the Moses engines using MMM is explained. Then an explanation of how the raw MT output was obtained, processed and evaluated will be given before presenting the results and drawing a conclusion.

Overview

A study carried out by Machado and Fontes (2011) for the Directorate General for Translation at the European Commission forms the basis for the methods adopted in the experiments. The aim was to explore integrating a personal SMT engine into a translator's workbench whereby the MT system was to be within a single computer with no connection to the Internet. It is trained with data that the user owns or has permission to use. The MT output is post-edited by the user to a level that Taus (2014) defines as being comprehensible (i.e. the content is perfectly understandable), accurate (i.e. it communicates the ST meaning) and the style is good but probably not that of a L1 human translator. Punctuation should be correct and syntax and grammar should be normal as follows:

- The post-edited machine translation (PEMT) should be grammatically, semantically and syntactically correct.
- Key terminology should be correctly translated.
- No information should be accidentally added or omitted.
- Offensive, inappropriate or culturally unacceptable content should be edited.
- As much raw MT output as possible should be used.
- Basic rules of spelling, punctuation and hyphenation should apply.
- Formatting should be correct.

McElhany and Vasconellos (1988, p.147) warn that because editing is not rewriting corrections should be minimal.

To carry out this study, I installed Moses on a desktop computer using MT software MMM that claims to be user-friendly and able to be understood by users who are not computational linguists or computer scientists. Such users are referred to as 'mere mortals' (Machado and Fontes, 2014, p. 2).

A large parallel corpus is required for training Moses. TM is ideal for this because it produces aligned bi-texts that can be used with minimal changes. The Canadian Parliament's Hansard, which is bilingual, was the source of data for early work on SMT (Brown et al, 1988, p. 71.)

A data source created and often used to promote the progress of SMT development is the Europarl corpus produced from the European Parliament's multilingual proceedings, which are published on the EU website. Koehn (2005) arranged them into the corpus. He confirmed that it can be used freely (personal communication, 25 January 2016). It was chosen to simulate a TM for this project because it was used in the study made by Machado and Fontes (2011). When aligned with German it has 1,011,476 sentences.

I used MMM to build four MT systems with different amounts of data and tested them with a test document of 1000 isolated sentences extracted, together with their translations from the corpus.

Moses' developers suggest that by varying the tuning weights it is possible to tune the system for a particular language pair and corpus (Koehn, 2015, p.62). MMM facilitates some adjustments and the effect of these was studied using the largest training.

Before explaining how the experiments were conducted I will describe how I installed MMM and built the Moses engines.

.

Equipment, and software installation

MMM (Machado, and Fontes 2014, p.2) is intended to make the SMT system Moses available to many users and is distributed under a GNU General Public Licence (p.10). There is a very comprehensive MMM tutorial (Machado, and Fontes 2014) giving a step-by-step guide to SMT for newcomers like myself.

The tutorial recommends a PC with at least 8GB of RAM, an 8 core processor and 0.5 TB hard disk (p.14) but no less than 4 GB of RAM and a 2 core processor. I used a machine with 8 GB of ram, 4 processors but only a 148GB hard disk. There is a 'transfer-to-another-location' script that can be used to transfer training to another machine with a much lower specification for translating/decoding only. I tried this using a 1GB laptop but would not recommend it. It was able to complete the translation but it took hours rather than the minutes taken by the 8GB machine.

MMM consists of a series of scripts that automate installation, test file creation, training, translation and automatic evaluation or scoring. Following the tutorial, I installed Ubuntu on the computer, choosing 14.04(LTS)(64 bits), although MMM will also run on 12.04 (LTS).

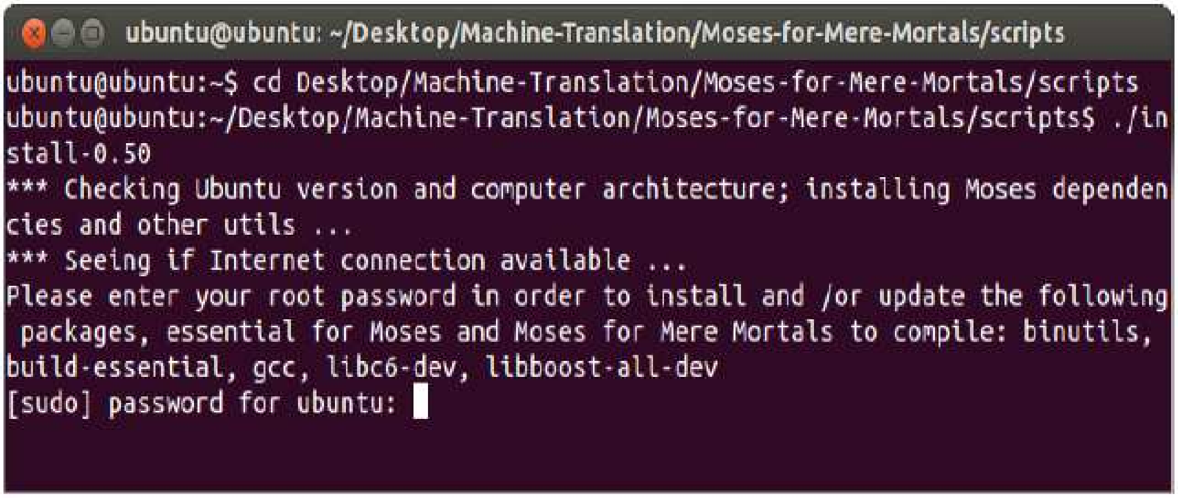
The next step was to download a zipped archive of MMM files and unpack them onto the computer.

The MMM tutorial explains how to prepare the corpus for training and build the system. Training the full Europarl corpus took 30 hours.

Although Ubuntu has a Graphical User Interface (GUI), I preferred the Command Line Interface (CLI) (see figure 2). The script 'Install', was run next to install all the files required onto the computer. MMM includes all the files necessary to run Moses but the script downloads, any Ubuntu files that are needed but not present in the computer from the Internet

With MMM installed running the 'Create' script completes installation of Moses.

There is a demo corpus that translates from English into Portuguese included with MMM for trying out Moses. I used this to experience preparing the corpora, extracting test data, translating and scoring before doing it with the German and English parts of the Europarl corpus.



```
ubuntu@ubuntu: ~/Desktop/Machine-Translation/Moses-for-Mere-Mortals/scripts
ubuntu@ubuntu:~$ cd Desktop/Machine-Translation/Moses-for-Mere-Mortals/scripts
ubuntu@ubuntu:~/Desktop/Machine-Translation/Moses-for-Mere-Mortals/scripts$ ./in
stall-0.50
*** Checking Ubuntu version and computer architecture; installing Moses dependen
cies and other utils ...
*** Seeing if Internet connection available ...
Please enter your root password in order to install and /or update the following
 packages, essential for Moses and Moses for Mere Mortals to compile: binutils,
build-essential, gcc, libc6-dev, libboost-all-dev
[sudo] password for ubuntu: █
```

Figure 1 Installing Ubuntu packages

Preparation of the corpora.

The 'Make-test-files' script was used to extract a 1000 segment test file from the Europarl corpus before using it for training.

Training

With MMM it is possible to build multiple translation engines on one computer. Where possible files generated by earlier ‘trainings’ are re-used. The training to be used for a particular translation is selected from the ‘translation’ script.

A monolingual TL corpus is used to build the language model. I used the English side of the bilingual corpus in all trainings. The aligned texts are placed in a folder named ‘corpora-for- training’ and the train script is run. When the training is complete a report is generated that is required by the ‘translation’ script to select the correct training.

Four basic trainings were built and tested. The first used the whole corpus. Then a second engine was built by splitting out the first 200,000 segments. This was repeated for 400,000 and 800,000 segments. The 1,000 segment test document was translated by each of the engines and Bleu scores obtained using the MMM ‘Score’ script. A sample of 50 segments from each translation was post-edited and evaluated by me.

Translation

The tests were divided into two parts. Before studying the difference in translation quality using the different sized corpora, the effect of the tuning weights was examined with the whole corpus.

With the German ST part of the test document in the ‘translation-in’ folder and the required data entered into the ‘Translate’ script, translation was initiated by running the script from the CLI.

According to Koehn (2015, p.62) a good phrase translation table is the key to good performance. He goes on to explain how some tuning can be done using the decoder. Significantly, the weighting of the four Moses models can be adjusted. They are combined by a log linear model (Koehn, 2010, p.137), which is well known in MT circles. The four models or features are:

- The **phrase translation** table contains English and German phrases that are good translations. A phrase in SMT is one or more contiguous words. It is not a grammatical unit.
- The **language model** contributes by keeping the output in fluent English.
- The **distortion model** permits the input sentence to be reordered but at a price: The translation costs more the more reordering there is.
- The **word penalty** prevents the translations from getting too long or too short.

There are three weights that can be adjusted in the 'Translation' script. These are Wl, Wd and Ww. They have default values of 1,1 and 0.

The tuning weights were adjusted in turn using the translation script.

With all the weights left at their default levels I produced the first translation. The reference translation was placed in the MMM 'reference-translation' folder and a Bleu score was obtained by running the 'score' script. I then post-edited 50 segments and evaluated the MT as explained below. This was repeated with Wd set to 0.5 and then 0.1. Then with Wd set back at 1, and with Wl set to 0.5 and then 0.1 further MTs were gathered and evaluated.

Similar experiments were conducted with Ww set to –3 and then 3 and with Minimum Bayes Risk (MBR) decoding (MBR decoding outputs the translation that is most similar to the most likely translation).

Having explained how the system was built and the MTs obtained, the methods used to evaluate the results will be described.

Evaluation

A total of 8 measurement points generated translations that were evaluated by both automatic and manual techniques.

Metrics

Machado and Fontes (2011, p.4) utilised the automatic evaluation metric Bleu (bilingual evaluation understudy), which compares the closeness of the MT to a professional translation relying on there being at least one good quality human reference translation available (Papineni et al, 2001, p.1). It is measured with an n-gram algorithm developed by IBM. The algorithm tabulates the number of n-grams in the test MT that are also present in the reference translation(s) and scores quality as a weighted sum of the counts of matching n-grams. In computing the n-gram overlap of the MT output and the reference translation the IBM algorithm penalises translations that are significantly longer or shorter than the reference. For computational linguists Bleu is a cheap quick language independent method of evaluation and correlates well with human techniques (Papineni et al, 2001).

In many cases this correlation has been shown to be correct (Doddington, 2002, p.138-145) and a study by Coughlin (2003, p.6) claims that Bleu correlates with the ranking of the TM and also 'provides a rough but reliable indication of the magnitude of the difference between the systems'. However, Callison-Burch et al (2006) take the view that higher Bleu scores do not necessarily indicate improved translation quality and focused manual evaluation may be preferable for some research projects. They conclude that for systems with similar translation structures Bleu is appropriate.

Manual Evaluation

Machado and Fontes (2011) had a team of translators performing human translations of a sample of segments even though human evaluations of MT output are extensive, expensive and take weeks or months to complete (Papineni, Roukos, Ward and Zhu, 2001, p.1). White (2003, p.213) points out that they are very subjective because there is no ‘right’ translation, as there is never any agreement on which is the best. Newmark (1982, p.140) is convinced that the perfect translation does not exist, but if it does Biguenet and Schulte (1989, p.12) are sure that it will never be found. Evaluators are always biased (White, 2003, p.219). For example, seeing a really bad segment might make the next one seem relatively better than it is and vice-versa. Another example is where a mistake such as a trivial software bug is forgiven. An evaluator may also become bored or tired, resulting in segments graded early in the cycle receiving a more favourable treatment to those graded later.

'Fluency' and 'adequacy' are commonly used for evaluating raw MT output. Two scores are combined, averaged, and written as a percentage. Machado and Fontes (2011, p.7) did not follow this method and use what they call 'real life conditions' by post-editing and classifying the effort required for each segment on a scale of 1 to 5.

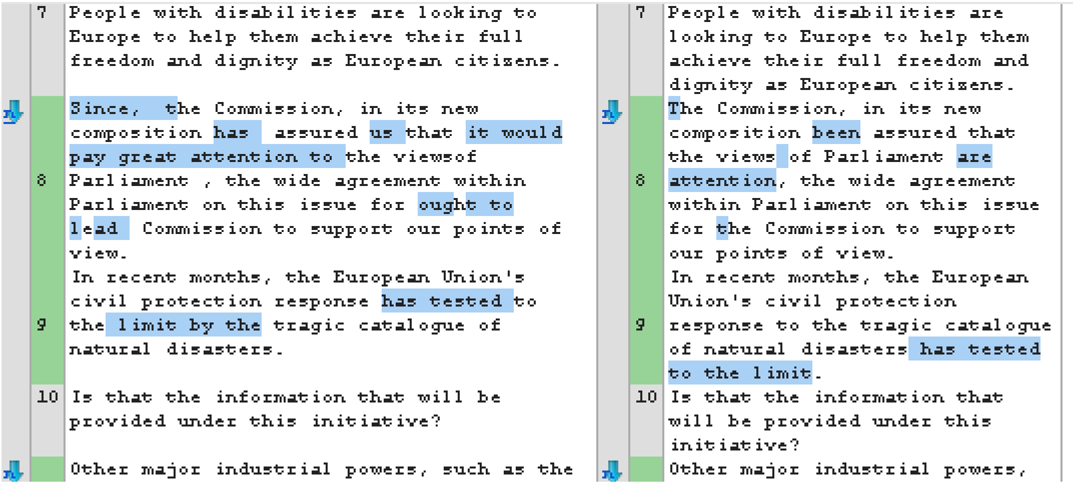
Their scale was adopted in this study:

1. Bad: Many changes for an acceptable translation; no time saved.
2. So So: Quite a number of changes, but some time saved.
3. Good: Few changes; time saved.
4. Very Good: Only minor changes, a lot of time saved.
5. Fully correct: Could be used without any change, even if I would still change it if it were my own translation.

Machado and Fontes do not mention whether or not time saved was measured but they do say that their objective was classifying segments by translation quality. Only a translation that can be used without change scores 5. A segment that is understandable and correct apart from one or maybe two errors receives a score of 4. One that should be translated from scratch scores 1.

Scores were recorded segment-by-segment on a spreadsheet and averaged for the fifty segments. Since I was the only post editor 50 different segments were post-edited for each MT. This avoided previous knowledge influencing the scoring and permitted a PEMT version of the test text to be gradually produced. For consistency with the Bleu scores the averages were divided by 5 to express them on a scale of 0 to 1.

There were eight measurement points in the first part of the experiments. A further four measurement points were made for the second part. These were for the 200000, 400000, 800000 and, for comparison, GT.



7	People with disabilities are looking to Europe to help them achieve their full freedom and dignity as European citizens.	7	People with disabilities are looking to Europe to help them achieve their full freedom and dignity as European citizens.
8	Since, the Commission, in its new composition has assured us that it would pay great attention to the viewsof Parliament , the wide agreement within Parliament on this issue for ought to lead Commission to support our points of view.	8	The Commission, in its new composition been assured that the views of Parliament are attention, the wide agreement within Parliament on this issue for the Commission to support our points of view.
9	In recent months, the European Union's civil protection response has tested to the limit by the tragic catalogue of natural disasters.	9	In recent months, the European Union's civil protection response to the tragic catalogue of natural disasters has tested to the limit.
10	Is that the information that will be provided under this initiative?	10	Is that the information that will be provided under this initiative?
	Other major industrial powers, such as the		Other major industrial powers,

Figure 2 from text-compare.com

For the second part a third method of evaluation was introduced. This was based on Translation Edit Rate (TER), as a less subjective technique to check the quality of assessment, which should be quantitative.

TER is defined as the lowest number of edits needed to change a MT segment so that it matches the reference, which is the post-edited segment, normalised by the length of the reference (Snover et al, 2006, p.3).

TER = number of edits / number of words in the PEMT sentence.

The raw MT was compared with the post-edited text using text-compare.com as shown in figure 3. The number of edits and the number of words in the PEMT were counted manually.

I termed this hter because of the human involvement. I used a spread sheet to obtain a score for each segment, which were then ranked on a scale of 1 to 5 as follows:

TER	Hter
0	5
0 to 0.25	4
0.25 to 0.5	3
0.5 to 0.75	2
>0.75	1

These segment scores were then averaged over the 50 segments to obtain an overall score.

Since the problems associated with using online MT systems are the rationale for investigating personal SMT, a comparison with the online MT engine GT was made. In order to compare the quality of the MMM trainings with GT, an MT was generated using GT and Bleu, human and hter scores were also obtained.

Results

Before seeing and discussing the results we will look at some sample translations that demonstrate the scoring levels. The source text (ST) is a segment from the test document and the reference translation (RT) is the corresponding English segment. The raw machine translation (MT) was produced by Moses. The MT was post-edited (PE) by me.

Starting with an example that scored 5

ST : *Wir müssen und können handeln.*
MT : We can and must act.
RT: We can and must take action.

I considered that this was a good translation based on the TAUS guidelines and did not require post-editing. The reference translations are given as an aid to non-German speaking readers. They are not necessarily better or worse than the MT or my PEMT.

The next example was given a score of 4. It only needed a few minor edits.

ST: *Wir halten es für unbedingt notwendig und nicht weniger dringlich, dass wir alle gemeinsam - und natürlich mit der völlig unabdingbaren Unterstützung dieses Parlaments - darauf hinwirken, dass dieses Recht der Petersberg-Aufgaben sofort zur Anwendung kommen kann, wenn diese Missionen ausgeführt werden.*

MT: We believe it is essential, and no less urgent, that amongst all of us - and with the completely indispensable cooperation of this Parliament - we start creating this law for Petersberg tasks to be applied if this mission

PE: We believe it is absolutely essential, and no less urgent, that between all of us - and with the completely indispensable cooperation of this Parliament - we start working towards this law for Petersberg tasks being immediately applied if these missions are carried out.

RT: We believe it is essential, and no less urgent, that amongst all of us - and with **t**he completely indispensable cooperation of this Parliament of course - we start creating this law for Petersberg tasks, which can be applied from the start of any mission.

A score of 3 was given to the following example. The MT cannot be understood

ST: *Die Vorstellungen des Vorsitzes im Umweltbereich klingen zwar gut, sollten aber in Resultate umgemünzt werden.*

MT: Which the presidency on the environment is sound, but results umgemünzt.

PE: The presidency's ideas on the environment sound good, but should be converted into results.

RT: The presidency's ideas in the environmental field sound good but should be translated into results.

Weight change	Average Human score	Average Human score /5	Bleu score
Default	3.78	0.756	0.5076
Distortion weight=0.5	3.58	0.716	0.5955
Distortion weight = 0.1	3.48	0.696	0.5912
Word penalty weight =3	2.36	0.472	0.342
Word penalty weight = -3	2.88	0.576	0.376
Language model weight =0.1	3.82	0.764	0.5076
Language model weight = 0.5	4.1	0.82	0.6203
Maximum Baye’s risk = 0	3.66	0.732	0.5948

Table 1 Results from first part of experiments.

The following MT does not require re-translating from scratch but it needs a lot of editing. It scored 2.

ST: *Vor Ihnen liegen zwei große Hindernisse, und zwar geht es darum, ob wir uns für die gegenseitige Anerkennung oder die Standardisierung entscheiden.*

MT: To two large obstacles, is whether we mutual recognition and the standardisation.

PE: Two large obstacles lie ahead of you. It is about whether we opt for mutual recognition or standardisation.

RT: Two big obstacles lie ahead of you. There is a problem of mutual recognition versus standardisation.

A score of 1 was given to the following MT because the ST had to be translated from scratch. My translation is more literal than the reference translation and *meine Vorredner* is plural.

ST: *Ich möchte meine Vorredner unterstützen.*
MT: To others.
PE: I would like to support the previous speakers.
RT: I would like to second what the previous speaker had to say.

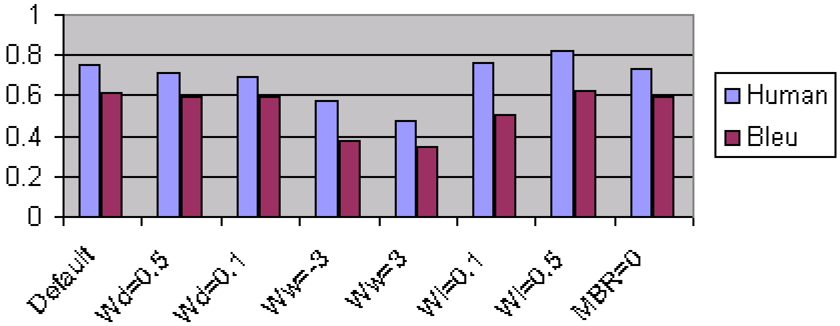


Figure 3 Results from first part of experiments

From table1 and figure 3 we can see that reducing the distortion weights (Wd) reduces the translation quality marginally.

For example the first segment in the test data is

ST: *Ich möchte meine Vorredner unterstützen*.

The MT with default weights is: 'To others'. With the distortion weight at 0.5 it is the same but with the distortion weight reduced to 0.1 it is 'I would support my' which is clearly different.

Varying the word penalty weight had a greater effect but reduced the quality for both increasing and reducing the weighting. Negative values should favour longer output and positive values should prevent short translations.

For the ST sentence *Wir haben dann abgestimmt.* [We then put it to a vote.. (my translation)]

With word penalty weight set to 0, 3 and -3 the MTs were:
We have voted.
We voted.
At the same time, we have to say that we will be able to put to the vote.

Surprisingly reducing the LM weighting increased the MT quality indicating that with the default weights this training favours a poorer translation if it is better English.

The following example illustrates this.

Auch hier möchte ich Sie darauf verweisen, daß das Verfahren beschleunigt werden muß.

[Also here would like I you thereon refer, that the process speeded up become must].

This MT with language model weight =1 has been favoured
‘You must realise that the process needs to be accelerated'.

With the language model weight = 0.5 the better translation- 'Also here I would like to remind you that this process needs to be accelerated' is produced. The problem is that *möchte ich Sie darauf verweisen* means ‘may I remind you’ but it is mis-translated in the corpus as ‘you must realise’ and given a low probability of being a translation by the phrase table. It is given a higher probability by the LM than the correct translation. Reducing the LM weighting diminishes this effect.

Size of training corpus	Average Human score	Average Human score/5	Bleu score	Average Hter score	Average Hter score/5
200000 segments	2.52	0.504	0.2385	2.4	0.48
400000 segments	2.86	0.572	0.308	2.76	0.552
800000 segments	3.64	0.728	0.31	3.68	0.736
Full corpus	3.77	0.754	0.6129	3.9	0.78
Google	3.76	0.752	0.31	4.085	0.817

Table 2 scores for different sized corpora and GT

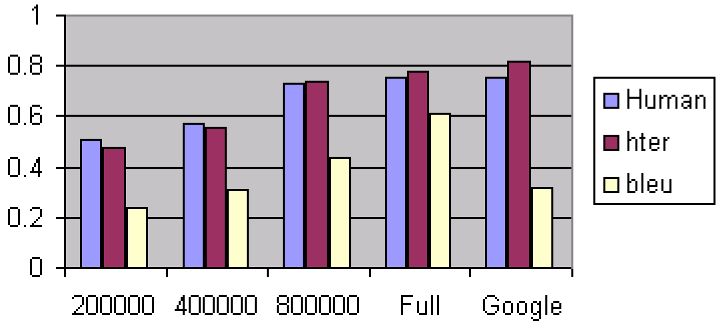


Figure 4 Results in chart form

As expected the quality of the MT output increases with the size of the training data as shown in table 2 and figure 4. The Bleu score trend follows the human and hter scores for the MMM trainings but the Google Bleu score is lower showing agreement with the notion that Bleu scores cannot be used to compare two MT systems with different architectures (Štajner et al, p.595).

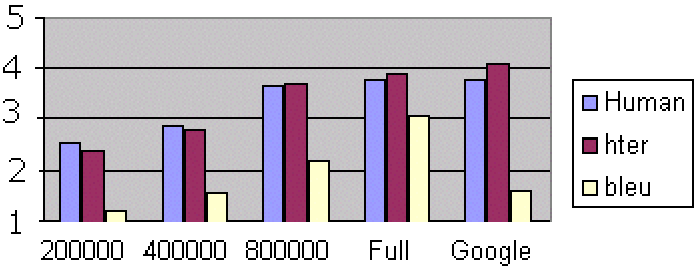


Figure 5 Results on 5 point base

In figure 5 the chart scale has been redrawn to reflect the five-point scale of the human and hter scores. For this 0.2 corresponds to a score of 1 and 0.4 is an average score of 2.

None of the trainings scored 1, the level for which post editing is not worthwhile, but equally none of them crossed the 4 threshold. Robinson (2012, p.38) discusses the use of Google translate to create a first translation draft. This sets a benchmark the equal of which should be the aim of a personal SMT engine.

Looking again at the first segment of the test data.

ST *Ich möchte meine Vorredner unterstützen.*
All of the MTs are poor and I would score them as 1

Size of training	MT
200k	I others.
400k	To others
800k	My previous
Full	To others

Whereas Google's MT I would score 5 following the TAUS guidelines even though strictly 'want to' should be edited to 'would like to'.

Google	I want to support the previous speakers.
PE	I would like to support the previous speakers.
Ref	I would like to second what the previous speaker had to say

In figure 6 the 8.5% improvement achieved for manual scoring with a language model weighting of 0.5 has been applied to all the trainings on the basis that they are the same language and genre. The full corpus gives a score of 0.8 or an average score of more than 4 and 800000 segments are needed to equal GT, which according to Champollion (2007, p.2) represents, for an average translator, producing 50,000 translation units a year, 16 years work. Additionally the freelancer may not have the rights to use this material especially the STs that belong to the author who may withhold permission to include them in the corpus.

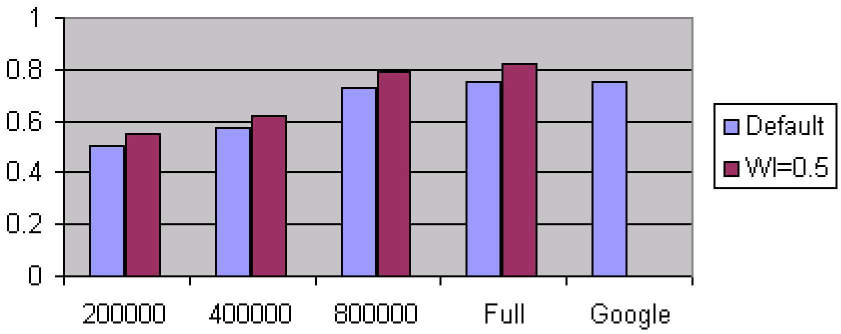


Figure 6 Corrected for Wl=0.5

Conclusion

This case study has successfully shown that it is possible to build a personalised SMT engine with MMM and that the quality of SMT output is directly proportional to the amount of training data available. In this instance for a performance equal to an online MT system the amount of data required was over 75% of the full Europarl corpus of 1.1 million segments. This is very high compared to the amount of material that an individual translator is able to produce, indicating that freelancers may struggle to find enough data to build an adequate system.

In addition to having enough material to build an SMT system with MMM a reasonably high degree of IT ability and knowledge is required or at least an interest in getting involved, even though it is aimed at translators rather than computational linguists and is free. Considering that there is a need to expand the amount of translation capacity available these results are disappointing for freelance translators.

Although this study is only a 'first glance' at using Moses as a personal MT engine it shows that SMT requires the very large amounts of data that are available to online translation engines. It was not carried out by a computational linguist/technologist but by a translator, which is important because now that translators have started to use MT as a tool to quickly produce a first draft, the translation community should take more interest in the development of MT tools. Somehow MT has to embrace and be embraced by TS. My observation from this study is that Pym's vicious circle is rooted in the fundamental techniques of SMT. The vast amount of data needed is far too much to be reasonably checked by humans but Moses generates its probabilities on what it sees in the training data and recycles the errors. Another source of errors observed is caused by the communicative nature of translations and the way that PBSMT relies on word alignment. Finding ways to improve the quality of MT with a limited size bi-text would help to provide post-editing and predictive tools for freelance translators. A first step might be to build an engine with real TM data in a specialised field and conducting experiments. Techniques such as hierarchical phrase tables might then permit data harvested from the Internet to be used but this would be a move away from baseline MMM requiring input from computer scientists.

References

Bowker, Lynne . 2005. Productivity vs Quality? A Pilot Study on the Impact of Translation Memory Systems. In *Localisation Focus* 4(1): 13–20

Brown, Peter. Cocke, John. Della Pietra, Stephen. Della Pietra, Vincent. Jelinek, Frederik. Mercer, Robert and Roossin, Paul. 1988. "A statistical approach to language translation." In *Proceedings, 12th International Conference on Computational Linguistics (COLING-88).*Budapest, Hungary, pp. 71-76.

Callison-Burch, Chris. Osborne, Miles. and Koehn, Philipp. 2006. Re-evaluating the role of Bleu in machine translation research. In *Proceedings EACL*, pp. 249-256 Retrieved from http://www.citeulike.org/user/JeremyKahn/article/2945969

Carson-Berndsen,Julie, Somers,harold, Vogel, Carl and Way, Andy.2010.Integrated language technology as part of the next generation of localisation in *The international journal of localisation* 8(1).

Champollion,Yves. 2007.The free, universal TM: are idealism and pragmatism compatible*? Technical seminar on copyright, intelectual property and translation tools Barcelone*.Retrieved from http://www.fit-europe.org/vault/barcelone/Champollion.pdf

Coughlin, Deborah.2003.Correlating automated and human assessments of machine translation quality. In *Proceedings of MT Summit IX*

Doddington,G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *In Human language technology : Notebook proceedings*, pp.128-132 San Diego

Garcia, Ignatius. 2011. Translating by post-editing: is it the way forward? in *Machine translation* vol 25 pp.217-237

Google. 2014.*Google terms of service.* Retrieved from https://www.google.com/policies/terms/

Klabunde, Achim. 2014.Cybersecurity in the era of freely available machine translation service in internet. Paper presented at MT@Work - Public Service Redesigned? Retrieved from https://scic.ec.europa.eu/streaming/index.php?es=2&sessionno=f4661398cb1a3abd3ffe58600bf11322v

Koehn,Philipp. 2005. *Europarl: a parallel corpus for statistical machine translation.*MT summit 2005. Retrieved from http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl.pdf

Koehn,Philipp. 2010. *Statistical machine translation.* Cambridge: Cambridge University Press

Koehn,Philipp. 2015. *Moses statistical machine translation system user manual and code guide.* Retrieved from http://www.statmt.org/moses/manual/manual.pdf

Machado, Maria and Fontes, Hilario. 2011. *Machine translation: case study – English into Portuguese — evaluation of Moses in dgt Portuguese language department using Moses for mere mortals* Retrieved from http://ec.europa.eu/translation/portuguese/magazine/documents/folha37_moses_en.pdf

Machado, Maria and Fontes, Hilario. 2014.*Moses for Mere Mortals tutorial: A machine translation chain for the real world.* Retrieved from https://github.com/jladcr/Moses-for-Mere-Mortals/blob/dfbfe799ebee1e1e0a3fa370fb4c6050511d5b0c/Tutorial.pdf

McElhaney,Terrence And Vasconellos,Muriel. 1988. The translator and the postediting experience. In Vasconellos,M (ed). *Technology as translation strategy.* Amsterdam: John Benjamins

Papineni, Kishore. Roukos, Salim. Ward, Todd and Zhu, Wei-Jing. 2001. *Bleu: a Method for automatic evaluation of machine translation IBM Research Report* Retrieved from http://www1.cs.columbia.edu/nlp/sgd/bleu.pdf

Pym.Anthony. 2010.*Exploring translation theories*, Abingdon: Routledge

Pym.Anthony. 2012. Translation skill-sets in a machine-translation age *Journal des traducteurs / Translators' Journa*l Volume 58(3) , pages 487-503.

Robinson, Douglas. 2012.*Becoming a translator.* 3rd ed London: Routledge

Snover, Matthew, Dorr,Bonnie.Schwartz,Richard. Micciulla,Linea and Makhoul.John. 2006. A study of translation edit rate with targeted human n-notation. In *Proceedings of association for machine translation in the americas* Retrieved from https://www.cs.umd.edu/~snover/pub/wsmt09/terp_wsmt09.pdf

Štajner, Sanja, Querido, Andreia, Rendeiro, Nuno, Rodrigues, Jo˜ao and Branco, António. 2016. Use of Domain-Specific Language Resources in Machine Translation in *Proceedings of the International Conference on Language Resources and Evaluation (LREC) 2016*. Portoroz, Slovenia, May 25-27

Taus. 2010. *MT post-editing guidelines* retrieved from https://www.taus.net/academy/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines

White, John. 2003. How to evaluate machine translation. In Somers, Harold.(ed). *Computers and translation. A translator's guide*. Philadelphia, PA, USA: John Benjamins Publishing Company