Integrating MT at Swiss Post's Language Service: preliminary results

Pierrette Bouillon, Sabrina Girletti
FTI/TIM, University of Geneva
Boulevard du Pont-d'Arve 40
1211 Geneva, Switzerland
Pierrette.Bouillon@unige.ch
Sabrina.Girletti@unige.ch

Paula Estrella, Jonathan Mutal
FaMaF, University of Córdoba
Av. Medina Allende s/n
X5000HUA Córdoba, Argentina
pestrella@famaf.unc.edu.ar
jdm0113@famaf.unc.edu.ar

Martina Bellodi, Beatrice Bircher
Swiss Post Ltd
Wankdorfallee 4
3030 Bern, Switzerland
martina.bellodi@post.ch
beatrice.bircher@post.ch

Abstract

This paper presents the preliminary results of an ongoing academia-industry collaboration that aims to integrate MT into the workflow of Swiss Post's Language Service. We describe the evaluations carried out to select an MT tool (commercial or open-source) and assess the suitability of machine translation for post-editing in Swiss Post's various subject areas and language pairs. The goal of this first phase is to provide recommendations with regard to the tool, language pair and most suitable domain for implementing MT.

1 Introduction

Nowadays, the production environments of many companies incorporate MT for various reasons: it might be upon request of a client, an initiative to add new services to a company's assets or an attempt to cut costs and shorten delivery times. The technology can be developed by a third party or in-house, each solution having its own pros and cons.

Swiss Post's Language Service would like to integrate MT in their workflow in different contexts, ranging from gisting to professional post-editing, thereby allowing for reduced turnaround times. Hence, in collaboration with the University of Geneva and one of its partners, the University of Córdoba, a preliminary study was carried out to 1) select an MT engine (open source or commercial) and 2) determine the language pairs and subject areas for which MT would be most suitable. In particular, we focused on assessing the potential suitability of MT sentences for professional post-editing.

The source data used to train and test the different systems for the various language pairs are almost parallel, making it possible to compare results across less-studied pairs. In addition, when designing our experimental setting, we chose to put the focus on users, namely Swiss Post's professional translators, providing them with specific training before involving them in the evaluation process. We are convinced that when reorganizing the traditional workflow of professional translators, it is important to give them an active role in the change in order to foster acceptance and avoid biased evaluation due to reluctant MT users.

The paper is structured as follows: we first describe the available data for the various languages and subject areas (Section 2), then explain how we selected the MT engine (Section 3). We then present how the suitability of the MT for PE was assessed by Swiss Post's in-house translators (Section 4) and discuss the results (Section 4.4), before concluding (Section 5).

Pérez-Ortiz, Sánchez-Martínez, Esplà-Gomis, Popović, Rico, Martins, Van den Bogaert, Forcada (eds.)
Proceedings of the 21st Annual Conference of the European Association for Machine Translation, p. 281–286
Alacant, Spain, May 2018.

2 Data and subject areas

Swiss Post's Language Service primarily translates texts from DE(CH) into FR(CH), IT(CH) and EN(UK). The Service has diverse activities, with specific translation memories (TMs) available in different subject areas: vocational training (denoted *Modulo*), financial services (*PF*), process manuals (*PN*), and annual report (denoted *GB*). In addition, there is a big "master" TM (denoted *MTM*) which includes all the specific TMs, plus additional material. The data are almost parallel across language pairs, meaning that at least 65% of source sentences are shared as training data[1]. Since the volume of translated material is significantly lower for DE-EN, we decided to only consider the "annual report" (*GB*) domain for this language pair. Details on amount of data are shown in Table 1.

TMs	**DE-FR**	**DE-IT**	**DE-EN**
Modulo	99,612	107,128	–
PF	129,694	122,568	–
PN	23,131	23,447	–
GB	38,580	37,721	32,857
MTM	2,558,148	1,929,530	417,817

Table 1: Number of translation units in TMs, per language pair.

The language pairs involved in this project are quite challenging, as they involve highly inflected languages (German, French and Italian). Furthermore, language pairs such as DE-IT and DE-FR are underrepresented in the vast literature on MT, as most of the results deal with English (either as the source or target).

3 MT system selection

3.1 Solutions considered

The first part of the study was devoted to a comparative evaluation between two phrase-based MT engines: the open-source toolkit Moses (Koehn et al., 2007) and the commercial online platform offered by Microsoft (Translator Hub, MTH[2]).

These solutions are common options for a company willing to experiment with MT; one is a third-party platform – which only requires uploading data (and then paying for the deployment and employment of the system) – while the other is an in-house solution, which, on the one hand, allows the entire process to be fully controlled, but on the other hand, requires technical knowledge and computing resources.

3.2 Engine training and evaluation

We followed the training process (corpus tokenization, language and translation model training, tuning and testing on a disjoint set from training) using the tools provided by Moses and MTH[3]. After some experimenting, language models for Moses were trained using KenLM (Heafield, 2011) on 4-grams. For models created in MTH, additional preprocessing was needed before building systems as data had to be anonymized for confidentiality. Therefore, named entities, numbers (belonging to phone numbers, amounts, accounts, etc.), urls and emails were replaced by placeholders in training and test data.

Since there are specific TMs for each subject area and language pair, we tried different combinations in order to obtain the highest automatic scores. Using each specific TM individually (*PN*, *Modulo*, *PF*, *GB*) resulted in a small-sized training set leading to poor automatic scores, so we decided to perform two incremental rounds of training:

- Round 1 - using all TMs together as a mixed training set: in this case we tested them on the different domains to explore how the system performed. Both Moses and MTH models were trained for DE-IT/FR.[4]

- Round 2 - using only the *MTM*: in this case we did not train models in MTH, as previous tests had indicated that the results with Moses were better and we could therefore save on the cost of anonymizing the data.

Models were evaluated automatically using standard metrics BLEU (Papineni et al., 2002)[5] and Word Error Rate (WER), as well as internal human evaluations. Four different test sets, one per

[1] Between DE-FR and DE-IT. The percentage is lower for pairs with EN, as this corpus is significantly smaller than the others.

[2] https://www.microsoft.com/en-us/translator/hub.aspx

[3] For training processes, see:
http://www.statmt.org/moses/?n=Moses.Baseline
https://hubtest.microsofttranslator-int.com/Help/Download/Microsoft%20Translator%20Hub%20User%20Guide.pdf

[4] The DE-EN pair was added to the study in a second phase.

[5] Although MTH provides BLEU scores after training, we report BLEU scores calculated using the script available at ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13.pl.

domain, were used to test the behavior of the engine when confronted with new data (not included in the training corpus). Amount of testing data is shown in Table 2.

Test set	**DE-FR**	**DE-IT**	**DE-EN**
PN	1736		–
Modulo	2034		–
PF	1919	2378	–
GB	1829	1718	704

Table 2: Number of translation units in test set per language and domain. For *Modulo* and *PN*, IT and FR shared exactly the same source sentences, while in the other domains, at least 58% of the corpus was shared. This percentage is lower with EN, since the related corpus was significantly smaller.

3.3 Results of MT engine evaluation

Results for Round 1 of training are shown in Tables 3 and 4: Moses outperforms MTH in all domains and better scores are obtained for *PN*. On the basis of these results, Round 2 of training was implemented; we only trained Moses on *MTM* to avoid having to anonymize data sets. Results improved for all domains (see Table 5).

	Moses		**MTH**	
Test set	WER	BLEU	WER	BLEU
PN	43.93	0.51	55.11	0.36
Modulo	45.94	0.46	60.17	0.31
PF	50.92	0.40	63.84	0.28
GB	58.49	0.34	71.91	0.23

Table 3: Results for DE-FR on mixed training set (all TMs).

	Moses		**MTH**	
Test set	WER	BLEU	WER	BLEU
PN	40.40	0.52	52.68	0.37
Modulo	44.16	0.46	55.55	0.35
PF	46.43	0.43	58.36	0.32
GB	51.94	0.40	62.66	0.31

Table 4: Results for DE-IT on mixed training set (all TMs).

We concluded that we could safely proceed to the human evaluation of suitability for PE (detailed in Section 4) with only Moses trained on *MTM*.

Test set	lang/pair	WER	BLEU
PN	DE-IT	33.01	0.6
	DE-FR	34.39	0.61
Modulo	DE-IT	40.96	0.5
	DE-FR	43.53	0.5
PF	DE-IT	43.07	0.48
	DE-FR	41.14	0.52
GB	DE-IT	47.41	0.45
	DE-FR	54.28	0.39
	DE-EN	34.48	0.62

Table 5: Results for DE-FR/IT/EN on MTM.

4 Human evaluation: suitability of MT for PE

4.1 Goal

The aim of the evaluation was to assess the potential suitability[6] of MT for post-editing in various language pairs and subject areas, from the perspective of Swiss Post's translators. We decided to let the translators assess the quality of the segments first, before involving them in a real post-editing task, in order to give them an idea of expected quality.

4.2 Test data

For the human evaluation, we used four specific test sets. We randomly selected a sample of 250 German sentences per subject area (1000 sentences in total) from the original test sets (described in Table 2), along with their respective target translations in FR, IT and EN. The test sets are completely parallel, meaning that we selected exactly the same 250 source sentences per subject area across the three language pairs. As in the previous evaluation, we only used the subject area "annual report" (*GB*) for the DE-EN pair. The automatic scores for these specific test sets are shown in Table 6.

4.3 Methodology

Eight in-house translators of the Language Service participated in the test team: three for DE-FR and DE-IT, and two for DE-EN. All translators in the test team had been working at Swiss Post's Language Service for at least 6 months, and had 1 to 19 years of translation experience. Before performing the evaluation task, the test team was given a one-day training course on MT and PE, involving both

[6] We also use "usability of MT for PE" as a synonym for "suitability".

Test set	lang/pair	WER	BLEU
PN	DE-IT	35.91	0.58
	DE-FR	35.20	0.59
Modulo	DE-IT	41.88	0.48
	DE-FR	47.52	0.46
PF	DE-IT	47.32	0.41
	DE-FR	47	0.43
GB	DE-IT	47.46	0.43
	DE-FR	58.77	0.34
	DE-EN	41.78	0.51

Table 6: BLEU and WER scores for test set (250 sentences), per domain and language pair.

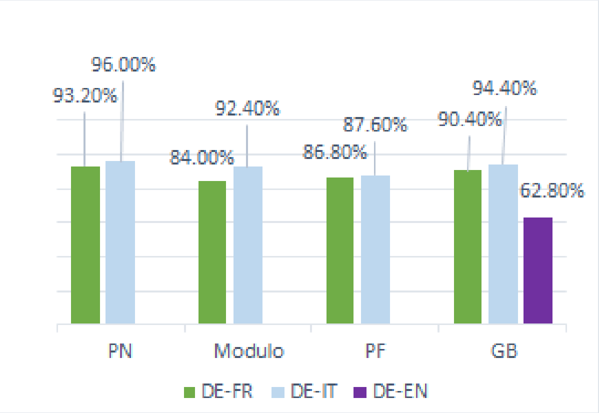

Figure 1: Percentage of machine translated sentences suitable for PE, per domain and language pair.

theory and practical exercises on MT engine training, evaluation and post-editing.

Since the purpose of this human evaluation was to assess the actual suitability of machine-translated sentences for subsequent post-editing by professional translators, we decided to use a customised metric. For each source sentence in the test sets, translators were presented with a raw machine translation and were requested to answer the following question: *"In a post-editing task, would you reuse this translation?"*, with possible answers being *"Yes, I would leave it as it is"* (denoted "Yes"), *"Yes, I would use it with some changes"* (denoted "YwC") and *"No, I would translate from scratch"* (denoted "No"). Since the evaluators were already familiar with the material being evaluated, we did not include any reference translation in our test. However, the translators were aware of the origin (that is, the subject area) of each segment, so that they could evaluate if the terminology used was appropriate.

We are aware that the "YwC" category is too broad, as it comprises all segments requiring minor changes or intensive post-editing, but in this preliminary evaluation we were mostly interested in finding out whether the translators would accept to post-edit the raw MT.

4.4 Results of human evaluation

Figure 1 summarises the results in terms of percentage of sentences suitable for PE, calculated as the sum of all "Yes" and "YwC" majority judgments[7], divided by the total number of sentences. In FR and IT, the results were very encouraging, with between 84% and 96% suitable sentences for each test set. The subject area *PN* obtained the best ratings, for both IT and FR; this result was also confirmed by automatic metrics (see Table 6). The second best domain was "annual report" (*GB*), with IT evaluators assessing a higher percentage of usable sentences than their FR and EN colleagues. However, this contradicts automatic scores, where GB seemed to be the subject area in which MT performed the worst. This calculation somewhat prejudices the scores of *GB* in EN, since there were only two evaluators and we only counted the sentences for which they agreed.

An Inter-Rater Reliability (IRR) analysis was performed to assess consistency among nominal ratings provided by the evaluators. Light's kappa (Light, 1971) and Cohen's kappa for DE-EN were used as an index of IRR. Figures are shown in Table 7.

	DE-FR	DE-IT	DE-EN
PN	0.341	0.549	–
Modulo	0.411	0.547	–
PF	0.412	0.519	–
GB	0.340	0.562	0.430

Table 7: Figures of Light's kappa (DE-FR/IT) and Cohen's kappa (DE-EN).

Overall, the results show moderate agreement among evaluators, with the exception of two domains (*PN* and *GB*) in DE-FR, where agreement is "fair" (Landis and Koch, 1977). Results are therefore more reliable for DE-IT.

Tables 8, 9 and 10 report detailed results per language pair. These results confirm that for DE-FR and depending on the domain, between 20-22% of the segments would not require post-editing at all ("Yes" column) and between 63.6-71.2% would

[7] Majority judgments are judgments on which at least two of the evaluators agree.

DE-FR			
ratings %	Yes	YwC	No
*PN**	22	71.2	5.2
*Modulo**	20.4	63.6	15.6
PF	22	64.8	13.2
*GB**	20	70.4	9.2

Table 8: Detail of ratings, per domain. For test sets *PN*, *Modulo* and *GB*, majority ratings could not be counted for, respectively, 2%, 0.4% and 0.4% of sentences.

DE-IT			
ratings %	Yes	YwC	No
*PN**	32.4	63.6	3.6
Modulo	31.6	60.8	7.6
*PF**	22.8	64.8	12
GB	26.8	67.6	5.6

Table 9: Detail of ratings, per domain. Majority ratings for the test sets *PN* and *PF* could not be counted for 0.4% of sentences.

require some post-editing, but could still be used. What remains to be studied is the effort it would take the translators to post-edit those segments in column "YwC" to convert them into a polished final translation. It is worth noting that the amount of segments that would be translated from scratch is minimal. Using the majority judgment, 2% of segments were scored in disagreement (i.e., they received three different scores).

Table 9 shows detailed results for DE-IT. The percentage of sentences in the "Yes" category was even higher than for the DE-FR language pair. In particular, the domain PN had the highest percentage of sentences usable without any modification ("Yes') and the lowest percentage of non-usable sentences ("No') overall.

For the DE-EN language pair, sentences could mostly be used with some changes. An equal percentage of "Yes" and "No" was also reported. However, it is worth noting that 28% of the sentences could not be counted. Since only two EN evaluators participated in the task, the majority judgment became a unanimous judgment, and we were not able to assess whether that third of the segments might be usable for post-editing. That is why, in Table 10, we also report minimal judgments, i.e. we count the times each nominal category received *at least* one score. When adding missing judgments to the count, more sentences are rejected and fewer sentences are accepted without any changes. However, in this particular case, we would need further analysis to confirm if DE-EN produces MT less suitable for post-editing, or if the evaluators are less inclined to use the raw MT output.

DE-EN			
ratings %	Yes	YwC	No
GB min.	14.8%	67.6%	17.6%
maj.*	9.2%	53.6%	9.2%

Table 10: Detail of minimal (min.) and majority (maj.) ratings, per domain. Since only two evaluators were involved in this task, majority ratings could not be counted for 28% of sentences.

Both human and automatic evaluations confirmed that "process manuals" (*PN*) is the best domain for the Language Service to begin implementing MT. We therefore focused on this domain to see what influences the subjective judgments of suitability, and further study if they correlate with objective factors (length of sentences, quality of raw translation).

As shown in Figure 2, translations assessed as "usable without modifications" ("Yes") are clearly shorter than the average length of source sentences in the corpus, while non-usable sentences ("No") are longer. The overall most chosen category, "YwC", comprises sentences that are generally longer than the average source sentence length.

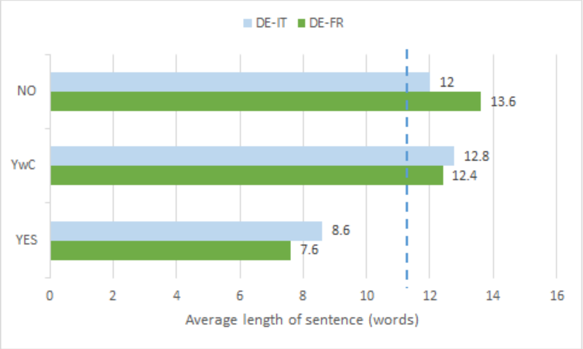

Figure 2: Average length of source sentences evaluated by IT and FR translators for the *PN* test set. Average sentence length in the *PN* source corpus is 11.37 words.

In table 11 we can see that WER scores also vary in line with suitability and that Light's kappa, calculated for each category, is inversely proportional to WER scores ("Yes" > "YwC" > "No").

Finally, we found that the amount of sentences that overlap in each category for both FR and IT is between 42%(IT) and 62% (FR) of "Yes" judgments, versus only between 31% (FR) and 44% (IT) for "No". These latest results are encouraging: they confirm that subjective judgments can be related to objective factors and that, in general, "Yes" judgments are very reliable, while "No" and "YwC" judgments seem to depend more on language, translators' choices and personal opinions.

lang. pair	metrics	**Yes**	**YwC**	**No**
DE-FR	%	22	71.2	5.2
	WER	20.40	37.34	65.16
	Kappa	0.462	0.282	0.235
DE-IT	%	32.4	63.6	3.6
	WER	22.99	42.68	71.29
	Kappa	0.64	0.514	0.339

Table 11: Detail of ratings (%) for *PN* compared to WER scores and Light's kappa (k) figures on the specific set of majority judgments for each category.

5 Conclusion and future work

We have presented the preliminary results of a project that aims to integrate MT into the workflow of Swiss Post's Language Service.

The first part of the study was devoted to choosing between the commercial Microsoft Translator Hub system and an in-house trained Moses solution, both trained using Language Service's material. We decided to proceed with the latter, trained on *MTM*, since automatic scores were systematically better with this system and training configuration. This allowed us to use just one system per language pair.

In the second part of the study, a human evaluation was carried out to assess the percentage of raw MT sentences perceived as suitable for professional post-editing. A sample of Swiss Post Language Service's professional translators was actively involved in this task. The outcomes of the evaluation were overall better for the subject area "process manuals" (*PN*). DE-IT evaluators assessed the highest percentage of usable sentences (with or without changes). More agreement among evaluators was also reported for this language pair. However, we sometimes found contradictions between human results and automatic scores, for instance in DE-EN, likely due to the fact that we only had two evaluators for this language pair. Furthermore, *GB* scored worse with automatic metrics, but was the second best subject area, according to human evaluation. Further investigation is required to discover the reasons behind this inconsistency between human ratings and automatic scores.

All in all, we consider our results to be satisfactory: a percentage of usable sentences ranging from 84% (DE-FR) to 96% (DE-IT) is a good threshold to start working with MT in a professional context. As for DE-EN, the 62.80% obtained suggests that in this case, raw MT output might be suitable, but to a lesser extent, so further work should be done in this direction.

In the next phase, we will carry out a productivity test with the translators, in order to determine if implementing MT into Language Service's workflow could actually be cost effective. These tests will first involve the highest scored domain (*PN*), since we believe that a gentle introduction to MT as new working tool is necessary to make the most of it. Finally, once translators are used to the new workflow, we would like to carry out a comparative evaluation of our PBMT system with the neural baseline we are currently training. This will allow us to compare both translators' productivity and satisfaction when using different MT architectures.

References

Heafield, Kenneth. 2011. KenLM: Faster and smaller language model queries. *Proceedings of the Sixth Workshop on Statistical Machine Translation* pp. 187–197, ACL.

Koehn, Philipp and Hoang, Hieu and Birch, Alexandra and Callison-Burch, Chris and Federico, Marcello and Bertoldi, Nicola and Cowan, Brooke and Shen, Wade and Moran, Christine and Zens, Richard and others. 2007. Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th ACL*, ACL.

Landis, J Richard and Koch, Gary G. 1977. The measurement of observer agreement for categorical data. *Biometrics*, JSTOR, pp. 159–174.

Light, Richard. 1971. Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological bulletin* vol. 76 nr. 5, American Psychological Association.

Papineni, Kishore and Roukos, Salim and Ward, Todd and Zhu, Wei-Jing. 2002. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th ACL*, pp. 311–318, ACL.