Moses SMT as an Aid to Translators in the Production Process

Falko Schaefer
SAP AG / Dietmar-Hopp-Allee
16, 69190 Walldorf, Germany
falko.schaefer@sap.com

Joeri Van de Walle
CrossLang N.V. / W. Wilson-
plein 7, Gent, Belgium
joeri@crosslang.com

Joachim Van den Bogaert
CrossLang N.V. / W. Wilson-
plein 7, Gent, Belgium
joachim@crosslang.com

Abstract

SAP has been heavily involved in the implementation and deployment of machine translation (MT) within the company since the early 1990s. In 2013, SAP initiated an extensive proof of concept project, based on the statistical MT system Moses (Koehn, et al., 2007), in collaboration with the external implementation partner CrossLang. The project focused on the use of Moses SMT as an aid to translators in the production process. This paper describes the outcome of the productivity evaluation for technical documents pertaining to SAP's Rapid Deployment Solutions (RDS), which was performed as part of this proof of concept project.

1 Background and Project Description

The use of machine translation at SAP dates back to the early 1990s. Originally the rule-based approach was deployed mainly for the translation of technical troubleshooting documents (SAP Notes), test cases, documentation, training materials, and as a gist translation tool for customer messages. MT systems used were METAL (German-English/English-German) and Logos (English-French mainly), followed by the next generation system Lucy LT (for these same languages). In 2012, SAP started experimenting with statistical machine translation (SMT). A prototype system was built at SAP Language Services (SLS) for the Chinese-to-English and English-to-Chinese language pairs. This prototype was based on the Moses SMT technology. In 2013, SLS initiated a more extensive proof of concept project, again based on Moses, in collaboration with the external implementation partner CrossLang. The project focused on the use of Moses SMT as an aid to translators in the production process. In that context CrossLang developed a plugin for SDL Trados Studio, thus enabling a seamless integration of Moses SMT into the SDL Trados Studio environment. MT suggestions were provided to translators during the proof of concept projects in addition to translation memory (TM) segments, which translators were free to accept, edit or discard just as they would TM matches. The overall timeline for the project was rather ambitious as all project phases (MT engine development, piloting, evaluation and engine improvement) had to be run between July and December 2013. In 2014, the SLS MT team will take additional steps to align machine translation landscapes and further extend the MT offering to various usage scenarios and more content types.

The proof of concept projects were carried out for two different content types: sap.com and RDS (Rapid Deployment Solutions) texts. While sap.com materials are typically texts used for SAP's official website, RDS texts are technical documents related to SAP's RDS product offering. Consequently the former content type can be classified as being of a more creative nature and thus more marketing-like than the latter, which is more technical by nature and hence more similar to documentation. The present paper will focus on the RDS content type.

The language scope of the proof of concept phase comprised the eight target languages Chinese, French, German, Italian, Japanese, Portuguese (Brazil), Russian and Spanish with source language English as well as the respective reverse language directions. However, the evaluations subject to this paper were carried out only for the target languages Chinese, French, German, and Russian.

For each language pair and content type Moses engines were built in three iterations:

- Iteration 1: Engines built with content type-specific data only (in-domain engines)

- Iteration 2: Engines built with a combination of content type-specific data and general SAP-related data to which domain adaptation techniques were applied (in-domain engines + domain adaptation)
- Iteration 3: Systems built in iteration 2 enhanced with natural language processing (NLP) components and techniques (in-domain engines + domain adaptation + NLP)

The size of the training data sets used for the relevant engines ranged from approximately 1 million to nearly 2 million tokens for sap.com and from roughly 2.2 million to 5.5 million tokens for the RDS content type.

While a total of more than 150 Moses engines were built throughout the project phase, not all of them could be run through the human evaluation rounds due to time and budget constraints. Instead, the best-performing systems for each content type and language pair were selected for human evaluation.

2 Evaluation Setup

In the proof of concept project, we looked at the machine translation output from different perspectives, which is reflected in the various types of evaluation that were performed: (i) engine development progress (automatic evaluation), (ii) translation quality (adequacy and fluency evaluation), (iii) translation productivity increase potential (productivity evaluation), and (iv) translation process (pilot projects).

The main goal of the automatic evaluations was to measure development progress. At the beginning of the project different test sets, each consisting of 1000 sentences, were extracted per content type and per language pair. Those test sets were held out of the training data and were used to score the engines after each development iteration. Three metrics were used for scoring: BLEU (Papineni, Roukos, Ward, & Zhu, 2002), METEOR (Banerjee & Lavie, 2005), and TER (Snover, Dorr, Schwartz, Micciulla, & Makhoul, 2006).

As automatic evaluation metrics are known to not always be reliable indicators of users' appreciation of the machine translation output quality, the automatic assessments were complemented with human judgments. With the adequacy & fluency evaluations, the focus was on the linguistic quality of the translations. With this evaluation we tried to answer the question 'how good is the translation?'. Two aspects of the translation were assessed: (i) in how far did the machine translation system succeed in transferring the meaning of the source sentence (adequacy), and (ii) in how far did the machine translation output respect the formal rules of the target languages (fluency). For both of these aspects, informants rated 400 machine translated sentences on a scale of 1 to 5, where 1 equals very poor performance and 5 excellent performance. For this evaluation, informants were linguists (professional and experienced translators). As with the automatic evaluations, sentences used for evaluation were kept apart from the training data.

With the productivity evaluation we tried to assess to what extent productivity increases might be obtained by using automated translation as an aid to speed up human translation. To evaluate this, informants were given a mix of (i) sentences pre-translated with MT, (ii) sentences without translation suggestion, and (iii) sentences with translations taken from the TM. The main reason for including the latter type of segments was to get an indication of the informants' possible bias against MT. Informants were asked to review and correct the translation of those sentences for which a translation was provided (MT or TM), and to come up with a translation from scratch for those sentences for which no translation was provided. In the background, the time informants spent on editing the translation output or translating the sentence was recorded. Recorded times were then used to calculate the average throughput for sentences in each of the categories (MT post-editing, TM match review, and translation from scratch). The sentences used in this evaluation were the same as those used for the adequacy and fluency evaluations and the informants taking part in this evaluation were provided by SAP's regular translation vendors.

When it comes to incorporating MT into the translation production process, a common concern is that the use of MT will negatively influence the quality of the translation output. The main objective of the pilot projects, finally, was to assess whether end-users of the translations would effectively notice quality differences between translations produced as the result of post-editing MT output and translations produced the traditional way. At the same time, the pilot projects served as a means to assess the complexity of integrating MT into the existing translation processes at SAP. To evaluate these aspects, MT was integrated into a real translation project, namely the translation of an update of existing contents for both content types in the pilot project. Sets of about 400 sentences per language were processed by SAP's regular translation vendors in two ways: once with a translation suggestion from MT and once as translation from scratch. The resulting translation variants were then compared and ranked by SAP employees in the target language countries.

3 Evaluation Results

Because of the space constraints for this paper, we will limit the discussion of the evaluation results to the adequacy/fluency evaluations and productivity evaluations of one particular content type, i.e. RDS. We discuss these results per evaluation type.

3.1 Adequacy/Fluency Evaluations

Table 1 shows the average ratings across informants for adequacy and fluency for all evaluated language pairs and the difference between the average adequacy and fluency ratings.

	En-to-De	En-to-Fr	En-to-Ru	En-to-Zh
Adequacy	4.11	4.16	3.54	3.89
Fluency	3.77	3.73	3.35	3.81
Difference	0.34	0.43	0.19	0.07

Table 1: RDS Adequacy/Fluency Results

Table 1 shows that the adequacy and fluency ratings vary per language, with the German and French output scoring best in terms of adequacy and the German and Chinese output scoring best as far as fluency is concerned. The lowest scores, both for adequacy and fluency, were observed for Russian.

The biggest difference between the adequacy and fluency rating was noted for the French output; the smallest difference for the Chinese MT translations.

3.2 Productivity Evaluations

Tables 2 through 5 show, for the four language pairs that were evaluated, the throughput per category per informant (in words per hour) and the productivity increase that is obtained by comparing the throughput for post-editing against that for translation.

	Translation	MT Post-editing	Full match revision	Productivity increase
Informant 1	288	358	458	24%
Informant 2	206	341	388	66%
Informant 3	347	599	1023	73%
Average	**280**	**433**	**623**	**54%**

Table 2: RDS Productivity Results En-to-De

For the English-to-German language pair, we observed an average productivity increase of 54% across informants for the RDS content type. A striking observation regarding the productivity evaluation for this content type is that, on average, 48% of the exact matches that were included in the evaluation set were changed by the informants. Informant 2 changed as much as 63% of the segments (i.e. 25 out of 40), which explains why his throughput for full match review is relatively lower than that of the other two informants.

	Translation	MT Post-editing	Full match revision	Productivity increase
Informant 1	531	906	1340	70%
Informant 2	451	628	617	39%
Informant 3	328	712	932	117%
Average	**437**	**749**	**963**	**76%**

Table 3: RDS Productivity Results En-to-Fr

For the English-to-French language pair, we observed an average productivity increase of 76% across informants. Looking at the results more closely, we found that there are considerable differences between the results of the different informants. For this content type, there is a difference of 78 percentage points between the increase noted for informant 3 (117%) and that noted for informant 2 (39%). We found evidence of the potential productivity gains in the fact that on average 40% of the segments with machine translation output remained unchanged. The lower increase noted with informant 2 might be explained by this informant having a more critical attitude towards MT. This becomes apparent when looking at the change rate for full match review segments. Informant 2 changed 58% (i.e. 23 out of 40) of the exact matches from TM as opposed to 50% for informant 1 and 35% for informant 3.

	Translation	MT Post-editing	Full match revision	Productivity increase
Informant 1	335	534	1541	59%
Informant 2	951	1833	4608	93%
Informant 3	296	443	889	50%
Average	**527**	**936**	**2346**	**67%**

Table 4: RDS Productivity Results En-to-Ru

For the English-to-Russian language pair, we observed an average productivity increase of 67% across informants. Although the average increase might be a little inflated by the high increase reported for informant 2, the fact that on average 41% of the sentences in the evaluation set was left unchanged by the informants, provides a good basis for explaining the observed increases. Interesting to see is that, compared to the languages already discussed, the Russian informants were less tempted to change full matches (on average only 28% were changed as opposed to 48% for both German and French).

	Translation	MT Post-editing	Full match revision	Productivity increase
Informant 1	266	333	453	25%
Informant 2	325	473	739	46%
Informant 3	264	312	420	18%
Average	**285**	**373**	**537**	**30%**

Table 5: RDS Productivity Results En-to-Zh

For the English-to-Chinese language pair, we observed an average productivity increase of 30% across informants. Again, we found that informants were very much inclined to change full match segments: on average 49% of the segments got changed. Whereas the change rate in the full match review category for the English-to-German and English-to-French languages pairs were found to be similar, the degree of change was a lot higher for the English-to-Chinese language pair (similarity score for full match review of 81.80) than for the language pairs discussed above (94.14 for English-to-German and 90.47 for English-to-French). This suggests that either the informants were very "picky" or that there was a problem with the quality of the TM. Further investigation revealed that the latter was the case.

4 Conclusions

The overall results of the quality evaluation as measured in the adequacy and fluency assessment appear rather encouraging across language pairs with a distribution of average scores between 3.35 and 4.16. There were, however, noticeable differences between individual languages with German and French scoring particularly well and Russian performing comparatively poorly in that part of the evaluation program. Apart from differing quality levels between the MT engines built for the various language pairs the fact that there is always an element of subjectivity involved in human quality judgments may serve as an explanation for this observation.

Besides the assessment of quality perception, another important question addressed in the evaluation rounds was obviously whether MT actually speeds up translation in the production process. As could be seen in the previous section, this question requires a differentiated answer depending on the target language, the main reasons for this being not only the varying quality levels of the engines for each language but also the fact that cultural aspects may impact the acceptance and hence perceived usefulness of machine translation as a translation aid. This became particularly apparent in the evaluation rounds for the target languages Russian and Chinese, where results of human quality and productivity evaluation were somewhat contradictory. However, this does not substantially affect the overall trend revealed by the productivity evaluation, which did prove clear productivity gains for all languages. This observation was confirmed by the translation vs. post-editing comparison in the pilot project evaluation (not discussed in this paper), which showed that the use of MT did not seem to have a negative impact on the quality of the final translations. As such translations produced with the help of MT were in no instance rated as being of lower quality than translations done from scratch. In fact quite the contrary was observed: For all languages a clear preference for translations resulting from MT plus post-editing could be established. This could be explained by the technical nature of RDS contents, where adequacy and fluency are considered more important than style and hence informants were less inclined to edit the MT output.

Finally it needs to be stressed that the evaluation results presented in this paper only reflect a snapshot of the quality of the engines built at the point in time the pilot and evaluation projects were conducted. Detailed system improvement activities are currently underway at SAP in order to further optimize MT engines and reach the defined quality levels for the various MT usage scenarios in the company.

References

Banerjee, S., & Lavie, A. (2005). METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. *ACL-2005: Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (pp. 65-72). Ann Arbor: University of Michigan.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., . . . Zens, R. (2007). Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (pp. 177-180). Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-j. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 311-318). Philadelphia, Pennsylvania: Association for Computational Linguistics.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, (pp. 223-231). Massachusetts.