Improving the Post-Editing Experience using Translation Recommendation: A User Study

[†]Yifan He **[†]Yanjun Ma** **[‡]Johann Roturier** **[†]Andy Way** **[†]Josef van Genabith**
[†]Centre for Next Generation Localisation, School of Computing, Dublin City University
[‡]Symantec Corporation Ireland
[†]{yhe,yma,away,josef}@computing.dcu.ie
[‡]johann_roturier@symantec.com

Abstract

We report findings from a user study with professional post-editors using a translation recommendation framework (He et al., 2010) to integrate Statistical Machine Translation (SMT) output with Translation Memory (TM) systems. The framework recommends SMT outputs to a TM user when it predicts that SMT outputs are more suitable for post-editing than the hits provided by the TM.

We analyze the effectiveness of the model as well as the reaction of potential users. Based on the performance statistics and the users' comments, we find that translation recommendation can reduce the workload of professional post-editors and improve the acceptance of MT in the localization industry.

1 Introduction

Recent years have witnessed rapid developments in statistical machine translation (SMT), with considerable improvements in translation quality. For certain language pairs and applications, automated translations are now beginning to be considered acceptable, especially in domains where abundant parallel corpora exist.

However, these advances are being adopted only slowly and somewhat reluctantly in professional localization and post-editing environments. Post-editors have long relied on translation memories (TMs) as the main technology to assist translation, and are understandably reluctant to give them up. There are several simple reasons for this: 1) TMs are useful as long as they are maintained; 2) TMs represent considerable effort and investment by a company or (even more so) an individual translator; 3) translators accept that the fuzzy match (Sikes, 2007) score used in TMs offers a good approximation of post-editing effort, which is useful for translation cost estimation; 4) translators are used to working with TMs and using something else could potentially have a negative impact on their productivity, at least in the short term, and 5) current SMT translation confidence estimation measures are not as robust as TM fuzzy match scores, and professional translators are thus not ready to replace fuzzy match scores with SMT internal quality measures.

One solution to promote the recent advances in statistical MT (such as (Koehn et al., 2003)) is to combine the strength of both worlds by integrating SMT with TMs. One of the main challenges of this integration is to establish a measure of confidence regarding the quality of MT output, similar to the TM fuzzy match scores for post-editors.

In our research we follow the approach of (He et al., 2010). Given that most post-editing work is (still) based on TM output, they propose to recommend MT outputs which are better (in terms of estimated post-editing effort) than TM hits to post-editors. In this framework, post-editors still work with the TM while benefiting from (better) SMT outputs; the assets in TMs are not wasted and TM fuzzy match scores can still be used to estimate (the upper bound of) post-editing labour.

(He et al., 2010) recast translation recommendation as a binary classification (rather than regression) problem using Support Vector Machines

(SVMs: (Cortes and Vapnik, 1995)) max-margin binary classifiers, perform Radial Basis Function (RBF) kernel parameter optimization to find the optimal meta-parameters for the classifier, employ posterior probability-based confidence estimation to support user-based tuning for precision and recall, experiment with feature sets involving MT-, TM- and system-independent features, and use automatic MT evaluation metrics to simulate post-editing effort. However, the evaluation in (He et al., 2010) suffers from lack of human-annotated data. Instead they use the TER automatic evaluation metric (Snover et al., 2006) to approximate human judgement. Despite the fact that the correlations between automatic evaluation metrics and human judgements are improving, professional post-editors are the ones that hold the final verdict over the quality of MT/TM integration. In order to draw grounded conclusions on the performance of the (He et al., 2010) recommendation framework, it is essential to conduct user studies to show whether or not systems developed using automatic evaluation metrics are confirmed by human judgements. Our experimental results support validation of the approach to approximate post-editing effort using an automatic evaluation metric (TER) in the translation recommendation model in (He et al. 2010): the model obtains more than 90% precision at above 75% recall against the judgements by professional human post-editing.

In this paper, we report the results of the human evaluation along with their behaviour and comments during the evaluation of a translation recommendation system similar to (He et al., 2010). We report findings on whether a high performance recommendation system trained on data annotated according to an automatic evaluation metric tallies with the judgements of professional post-editors.

The rest of the paper is organized as follows: we briefly introduce related research in Section 2, and review the classification model we adopted in Section 3. We describe the methodology of our user study in Section 4, and present an analysis of recommendation performance and user behaviour in Sections 5 and 6, respectively. Section 7 concludes and points out avenues for future research.

2 Related Work

The translation recommendation system we experiment with is an implementation of the translation recommendation model proposed in (He et al., 2010), which we review in more detail in Section 3.

Besides the translation recommendation model, there are several other models that try to combine the merits of TM and MT systems. The first strand is to design MT confidence estimation measures that are friendly to the TM environment, such as (Specia et al., 2009a) and (Specia et al., 2009b), both of which focus on improving confidence measures for MT, e.g. based on training regression models to perform confidence estimation on scores assigned by post-editors.

The second strand of research focuses on combining TM information into an SMT system, so that the SMT system can produce better translations when there is an exact or close match in the TM (Simard and Isabelle, 2009). This line of research is shown to help the performance of MT, but is less relevant to our task in this paper.

Moreover, (Koehn and Haddow, 2009) presents a post-editing environment using information from the phrase-based SMT system Moses (Koehn et al., 2007), instead of the fuzzy match information from TMs. Although all these approaches try to tackle the TM–MT integration task from different perspectives, we concentrate on evaluating the method of (He et al., 2010) in this paper.

The research presented in this paper focuses on aspects of a user study of post-editors working with MT and TMs. In this respect, it is related to (Guerberof, 2009), which compares the post-editing effort required for MT and TM outputs respectively, as well as (Tatsumi, 2009), which studies the correlation between automatic evaluation scores and post-editing effort. Our work differs in that this paper measures how the integration of TM and MT systems can help post-editors, not how post-editors perform using separate TM or MT systems.

3 The Translation Recommendation System

In this section we briefly review the translation recommendation system presented by (He et al., 2010). They use an SVM binary classifier to predict the rel-

ative quality of the SMT output to make a recommendation. The SVM classifier uses features from the SMT system, the TM and additional linguistic features to estimate whether the SMT output is better than the hit from the TM.

3.1 Problem Formulation

(He et al., 2010) treat translation recommendation as a binary classification between TM and SMT outputs, where the classifier recommends the output that is predicted to require less post-editing effort. They use automatic TER scores (Snover et al., 2006) as the measure for the required post-editing effort.

They label the training examples as in (1):

$$y = \begin{cases} +1 & \text{if } TER(\text{MT}) < TER(\text{TM}) \\ -1 & \text{if } TER(\text{MT}) \geq TER(\text{TM}) \end{cases} \quad (1)$$

Each instance is associated with a set of features from both the MT and TM outputs, which are discussed in more detail in Section 3.3.

3.2 Recommendation Confidence Estimation

In classical settings involving SVMs, confidence levels are represented as margins of binary predictions. However, these margins provide little insight for translation applications because the numbers are only meaningful when compared to each other. What is more preferable is a probabilistic confidence score (e.g. 90% confidence) which is better understood by post-editors and translators.

(He et al., 2010) use the techniques proposed by (Platt, 1999) and improved by (Lin et al., 2007) to obtain the posterior probability of a classification as a recommendation confidence score.

Platt's method estimates the posterior probability with a sigmoid function, as in (2):

$$Pr(y = 1|\mathbf{x}) \approx P_{A,B}(f) \equiv \frac{1}{1 + \exp(Af + B)} \quad (2)$$

where $f = f(\mathbf{x})$ is the decision function of the estimated SVM. A and B are parameters that minimize the cross-entropy error function F on the training data, as in (3):

$$\min_{z=(A,B)} F(z) = -\sum_{i=1}^{l} (t_i \log(p_i) + (1 - t_i) \log(1 - p_i)),$$
$$\text{where } p_i = P_{A,B}(f_i), \text{and } t_i = \begin{cases} \frac{N_+ + 1}{N_+ + 2} \text{ if } y_i = +1 \\ \frac{1}{N_- + 2} \text{ if } y_i = -1 \end{cases} \quad (3)$$

where $z = (A, B)$ is a parameter setting, and N_+ and N_- are the numbers of observed positive and negative examples, respectively, for the label y_i. These numbers are obtained using an internal cross-validation on the training set.

3.3 The Feature Set

(He et al., 2010) use three types of features in classification: the MT system features, the TM feature and system-independent features.

3.3.1 The MT System Features

These features include those typically used in SMT, namely the phrase-translation model scores, the language model probability, the distance-based reordering score, the lexicalized reordering model scores, and the word penalty.

3.3.2 The TM Feature

The TM feature is the fuzzy match (Sikes, 2007) cost of the TM hit. The calculation of fuzzy match score itself is one of the core technologies in TM systems and varies among different vendors. (He et al., 2010) compute fuzzy match cost as the minimum Edit Distance (Levenshtein, 1966) between the source and TM entry, normalized by the length of the source as in (4), as most of the current implementations are based on edit distance while allowing some additional flexible matching.

$$h_{\text{FM}}(t) = \min_{e} \frac{\text{EditDistance}(s, e)}{\text{Len}(s)} \quad (4)$$

where s is the source side of t, the sentence to translate, and e is the source side of an entry in the TM. For fuzzy match scores F, this fuzzy match cost h_{fm} roughly corresponds to $1 - F$.

3.3.3 System-Independent Features

(He et al., 2010) use several features that are independent of the translation system, which are useful when a third-party translation service is used or the MT system is simply treated as a black-box:

- Source-Side Language Model Score and Perplexity
- Target-Side Language Model Perplexity
- The Pseudo-Source Fuzzy Match Score: they translate the output back to obtain a pseudo source sentence. They compute the fuzzy match score between the original source sentence and this pseudo-source
- The IBM Model 1 (Brown et al., 1993) scores in both directions

4 Evaluation Methodology

We conduct a human evaluation on TM–MT integration with professional post-editors. In this section we introduce the evaluation data we use, the post-editors, the evaluation environment and the questionnaire which we give to the post-editors after they have completed the evaluation.

4.1 Data

Our raw data set is an English–French translation memory which consists of 51K sentence pairs of technical translation from Symantec. We randomly selected 43K to train an SMT system and used this system to translate the English side of the remaining 8K sentence pairs as recommendation candidates. We train SVM translation recommendation models with 4-fold cross validation on these 8K sentence pairs, and randomly select 300 from the cross validation test sets for human evaluation.

More specifically, for the SMT system, we use a standard log-linear PB-SMT model (Och and Ney, 2002): GIZA++ implementation of IBM word alignment model 4, the refinement and phrase-extraction heuristics described in (Koehn et al., 2003), minimum-error-rate training (Och, 2003), a 5-gram language model with Kneser-Ney smoothing (Kneser and Ney, 1995) trained with SRILM (Stolcke, 2002) on the target side of the training data, and Moses (Koehn et al., 2007) to decode.

For the translation recommendation model, we output a confidence level using the method in Section 3.2 and all the features in Section 3.3.

4.2 The Post-editors

Five professional post-editors help us to complete this study. Four of them are full-time post-editors, and one is a part-time post-editor. All of the editors are hired through the localization vendors of the IT security company and have experience on post-editing machine-generated segments (including TM, Rule-based MT or Statistical MT).

4.3 The Evaluation Environment

We design an evaluation environment to present the 300 English segments translated into French using the TM and MT systems to the post-editors. The environment is a web application developed in Python with the Django framework.[1]

Each post-editor is given a username and password to log in the system. After login, there is only one English segment together with its two French translations (from TM and MT) shown on each page. The two French translations are shuffled randomly: so translation 1 can either be output from MT or TM, and depending on the choice in 1, the same for translation 2the remaining option is provided as translation 2. In a production setting, we would present the recommended translation more favorably than the other, but as this experiment tries to evaluate the performance of the TM/MT integration technique, we need to keep it blind. A snapshot of the interface is shown in Figure 1.

The post-editors' operations in the system are recorded with a time stamp in the database, which allows us to analyze the time they spend on each segment. The system allows the users to log in and out of the environment so that their previous work is not lost. They are presented with the last segment they worked on after they log in again.

Each post-editor is provided with an introduction to the task before the experiment begins. Note that the post-editors are asked to choose the sentence that is most suitable for post-editing (which is also emphasized in the introduction to the task). The post-editors are told that even if a French translation does not fully translate the English segment, they may still select it because they would spend less time post-editing it into a grammatical French segment whose meaning would match the that of the English

[1] http://www.djangoproject.com

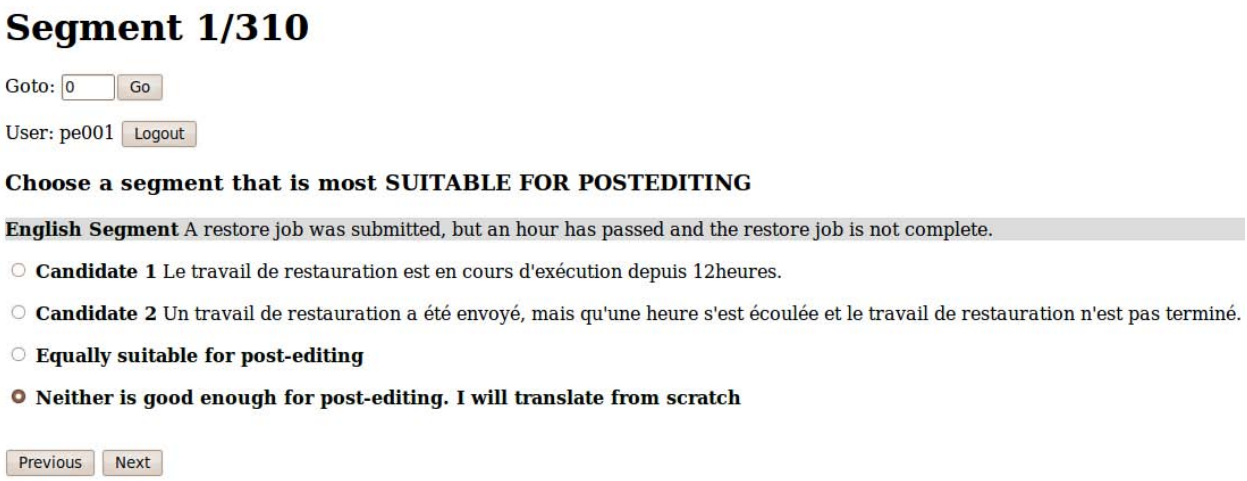

Figure 1: Interface of the Evaluation Environment

segment.

To control data quality and to measure intra-annotator correlation, we pre-select 10 segments from the 300 and make them appear twice in the environment. Therefore the post-editors are actually presented with 310 segments.

4.4 Questionnaire

After they have finished rating the 310 segments, the post-editors are presented with five questionnaire questions:

- Whether they are a full-time post-editor,
- If they are full-time post-editors, how long have they worked as a full-time post-editor,
- The number of words they have translated,
- Whether they have edited MT output professionally,
- What they think of MT (five choices: no idea, very useful, sometimes useful, not useful, and useless).

5 Analysis of Integration Performance

In this section we investigate the effectiveness of the translation recommendation model according to the judgements of professional post-editors. We also compare the results with the result on a gold standard approximated by TER scores to show whether it is valid to use automatic evaluation metric scores (to approximate post-editing effort) instead of human judgement in this task.

5.1 Precision and Recall of Translation Recommendation

We measure the precision and recall of the automatic translation recommendation, using the judgements of individual post-editors as a gold standard. We report the precision and recall numbers in Table 1. The precision can be further improved at the cost of recall, as we set the confidence threshold to 0.75 in Table 2. In these calculations, we discard the segments which the post-editors choose to translate from scratch, as translation recommendation cannot improve the post-editor's productivity in such cases, no matter what it recommends. When the post-editor chooses 'tie', we determine that the TM output should be preserved, in accordance with the conservative gold standard in (He et al., 2010).

Table 1: Precision and Recall of Recommendation, Individual Post-editors, confidence = 0.5

Post-Editor ID	Precision	Recall
PE01	0.8812	0.9223
PE02	0.9315	0.9315
PE03	0.8945	0.9138
PE04	0.9123	0.9369
PE05	0.8734	0.9409

In Table 1, the automatic recommendation obtains over 0.9 recall according to all post-editors. The pre-

Table 2: Precision and Recall of Recommendation, Individual Post-editors, confidence = 0.75

Post-Editor ID	Precision	Recall
PE01	0.9379	0.7824
PE02	0.9643	0.7621
PE03	0.9415	0.7629
PE04	0.9500	0.7703
PE05	0.9153	0.7864

cision of recommendation is always above 0.87. Table 2 shows that when the post-editors require more recommendation confidence, the translation recommendation can always obtain 0.9 precision at the cost of reducing recall. With these results on recommendation precision, there is a rather strong guarantee that the integrated MT-TM system will not waste the assets in the TM system and will not change the upperbound of related cost estimation, even at the sentence level, because the recommended SMT outputs are, in fact, more suitable for post-editing from the post-editors' perspective.

5.2 Precision and Recall on Consensus Preferences

The localization industry might expect even stronger confidence in the recommendation, so we measure recommendation precision on the segments where there is a consensus among the post-editors that the MT output should be used to post-edit.

To reflect consensus, we first discard the segments which the majority of the post-editors (more than 3 in this experiment) choose to post-edit from scratch. For the rest of the segments, we consider that MT output should be recommended, if N post-editors prefer to post-edit the MT output. Otherwise, we consider that the TM output should be recommended.

We report the precision and recall numbers on a series of confidence thresholds for $N = 3$ and $N = 4$ post-editors in Tables 3 and 4, respectively.

Table 3 shows that if we consider the consensus among 3 post-editors, precision is still high. Besides the capability of the recommendation system, there is also the reason that there are actually a larger number of segments to be recommended when we consider the consensus among 3 post-editors, rather than 1 post-editor. When we analyze Table 4, precision begins to drop. The reason for this is that this is an inherently more difficult task, and that the post-editor PE01 chooses to edit a larger number of segments from scratch (cf. Section 6.1), which renders it more difficult for the remaining post-editors to reach a consensus for $N = 4$.

Table 3: Precision and Recall of Recommendation, Consensus Preferences of $N = 3$ Post-Editors

Threshold	Precision	Recall
0.5	0.9110	0.9348
0.6	0.9412	0.9043
0.7	0.9606	0.8478
0.8	0.9689	0.6783
0.85	0.9695	0.5522

Table 4: Precision and Recall of Recommendation, Consensus Preferences of $N = 4$ Post-Editors

Threshold	Precision	Recall
0.5	0.8263	0.9420
0.6	0.8507	0.9082
0.7	0.8768	0.8599
0.8	0.8944	0.6957
0.85	0.8931	0.5652

5.3 The TER score and the Preference of Post-Editors

We measure the TER score of the TM and MT outputs, and sort them according to the post-editors' preferences in Table 5. The TER score is an edit-distance based metric that calculates the number of insertions, deletions, substitutions and shifts required to transform an MT output to a reference sentence, and is therefore expected to be a reasonable automatic metric to approximate post-editing effort. We report the results in Table 5, where the scores are averaged among the five post-editors.

Table 5: TER Scores Sorted by Preference

	TM	MT	Tie	Scratch
TM Output	25.00	57.37	19.16	70.33
MT Output	31.85	25.90	20.93	41.74

In Table 5, TER scores are shown to correlate well with post-editors' preferences: when the post-editor prefers MT, the MT output obtains a lower TER score, and vice versa. This validates the method

of (He et al., 2010), where the TER score is used to generate a gold standard for the translation recommendation system. The TER scores also demonstrate that the sentences which the users would translate from scratch are more difficult to translate in nature than the rest, as is shown by the deterioration of more than 15 TER points compared to the MT-output.

5.4 Comparison with a TER-Approximated Gold Standard

We present the precision numbers at recommendation confidence [0.5, 0.85] in Figure 2. Series PE01 – PE05 use the judgement of the corresponding post-editor as the gold standard; series CONSENSUS_3 and CONSENSUS_4 use the consensus of 3 or 4 post-editors as the gold standard; series TER uses the gold standard approximated by TER scores, as in (1). By presenting results on human-annotated and metric-approximated gold standards head-to-head, we are able to see the relationship between these gold standards.

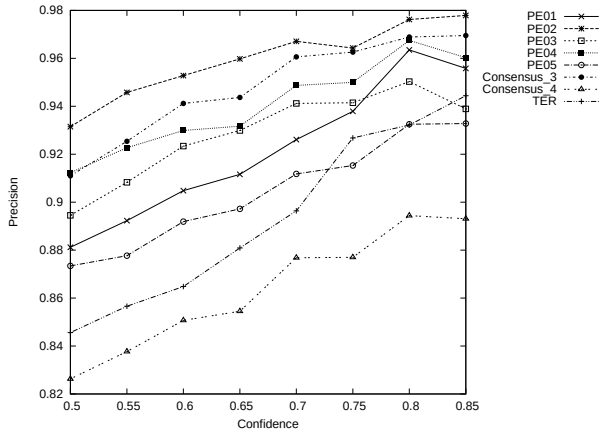

Figure 2: Recommendation Precision According to Human-Annotated and TER-Approximated Gold Standards

In Figure 2, we find that although the post-editors have different preferences regarding MT and TM outputs (i.e. some reuse MT outputs more than others), the trend of precision on the variation of confidence levels remains similar among the post-editors, and also applies to the TER-approximated gold standard. This again validates the approach of (He et al., 2010), which uses TER scores to approximate human judgements to prepare the training data and perform evaluation. Note that when calculating precision, the denominator is the total number of segments recommended by the recommendation model, no matter whether the post-editors have consensus judgements on them or not. If we limit the denominator to the number of segments where post-editors do reach a consensus judgement (on whether using the MT or the TM output), the precision will be 0.9641 for CONSENSUS_3 and 0.9848 for CONSENSUS_4.

6 Analysis of User Behaviour

Besides the performance of recommendation, we are also interested in the users' reaction to the TM-MT scheme using this system, as well as what they think about the TM and MT technologies. We report statistics of their behaviour along with their ideas and comments on TM and MT.

6.1 Experience of Post-Editors

We list the years of experience as translators of the post-editors along with the number of sentences they prefer to translate from scratch in our experiment in Table 6, because the latter is an indication of the willingness to reuse a computer-generated translation. We also present the number of MT outputs (out of 300) selected by post-editors to work on.

Table 6: Participants' Experience and Preference

Post-Editor ID	Years	Scratch	MT
PE01	5	59	193
PE02	3	11	248
PE03	12	22	232
PE04	8	33	222
PE05	part-time	23	220

The results show that the willingness to reuse automatic outputs varies considerably among post-editors. PE01 is willing to translate one-fifth of the sentences from scratch in this experiment, which is more than five-times the number of PE02. This preference does not correlate well with the years of experience, suggesting that this is more related to the particular habits of post-editors, rather than to their experience in the industry. The result also shows that all the post-editors select more MT outputs to post-edit than the other options.

6.2 Inter-annotator Agreement

To gauge the validity of human evaluation results, we computed the inter-rater agreement measured by Fleiss' Kappa coefficient (Fleiss, 1981) which can assess the agreement between multiple raters as opposed to Cohen's Kappa coefficient (Cohen, 1960) which works with two raters.

Fleiss' Kappa coefficient for our five post-editors is 0.464 ± 0.024, indicating a moderate agreement. We also obtained Fleiss' Kappa coefficient for each category as shown in Table 7. From this table, we can observe moderate agreements among post-editors in selecting TM or MT output as the most suitable for post-editing. There is also a moderate agreement in making their decision to translate from scratch. However, there is only a fair agreement in determining whether TM and MT outputs are equally good for post-editing ("Tie").

Table 7: Annotator agreement for each category

Category	Kappa
TM	0.519
MT	0.516
Tie	0.285
Scratch	0.426

6.3 Intra-annotator Agreement

We have ten duplicate samples in our evaluation intended to measure the level of intrinsic agreement for each post-editor. Both percentage of agreement and Cohen's Kappa are calculated as shown in Table 9. From this table, we can observe that all five post-editors achieved almost perfect intrinsic agreement, indicating that the evaluation results are highly reliable.

Table 8: Intra-annotator Agreement

Post-Editor ID	Agreement	Kappa
PE01	90%	0.87
PE02	100%	1.0
PE03	90%	0.87
PE04	80%	0.73
PE05	90%	0.87

6.4 Correlation between Sentence Length and Evaluation Time

Our evaluation interface is capable of logging the time spent by the post-editors in evaluating each sentence. One may expect that post-editors may spend more time in evaluating longer sentences and less time evaluating shorter sentences. We calculated Pearson's product moment correlation between the evaluation time and sentence length as shown in Table 9. The results appear to be inconclusive: we observe a high correlation between the evaluation time and sentence length for PE02 and PE05; however, for the other three post-editors, there is a low correlation. These inconclusive results can partly be attributed to the fact that we did not compel the post-editors to conduct their evaluation in one session. We expect to achieve more conclusive results in future work, which would happen in a real working post-editing environment.

Table 9: Pearson's Product Moment Correlation

Post-Editor ID	PMCC (r)	r-square
PE01	0.2246	0.0505
PE02	0.6957	0.4840
PE03	0.3916	0.1534
PE04	0.0746	0.0056
PE05	0.4907	0.2408
Average	0.2274	0.0517

6.5 Post-editors' Comments on MT and TM

We requested post-editors to comment on their attitude to MT and TM. In our questionnaire, all post-editors claim that they have post-edited MT outputs and think that MT is sometimes useful, which might represent the current state of MT penetration in the localization industry.

However, the more interesting comment comes from one of our post-editors in private communication. Although the post-editor does not know which of the two candidates we present in the evaluation interface is from the MT system, he claims after completing the evaluation that he has found that the TM outputs are more suitable for post-editing, although in fact every post-editor prefers MT outputs in the experiment (cf. Table 6).

This comment is revealing for two reasons. First

of all, the post-editor obviously mistakes MT outputs for TM outputs, which indicates that in this closed-domain setting mainly composed of simple short sentences, a state-of-the-art phrase-based SMT system is able to produce outputs that are not only correct on the word-to-word level, but also grammatically acceptable enough to be recognized as human translations in the TM, and therefore that the SMT output can be smoothly integrated into the TM environment.

Furthermore, the comment also shows how much the post-editors subconsciously trust the TM. This may be an explanation for the relatively low acceptance of MT technology in the localization industry, and demonstrates the need for TM–MT integration techniques such as translation recommendation.

7 Conclusions and Future Work

In this paper, we evaluate the effectiveness of translation recommendation (He et al., 2010) in the context of TM–MT integration with professional post-editors.

We find that a translation recommendation model trained on automatic evaluation metric scores can obtain a precision above 0.9 and a recall above 0.75 with proper thresholds according to each of the post-editors. The model shows precision above 0.8 when we evaluate against the consensus of post-editors. This supports validation of the method of (He et al., 2010) which uses automatic evaluation metrics to approximate actual post-editing effort.

From the analysis of user behaviour, we note that the users show consistency in their judgements according to both the inter-annotator agreement and the intra-annotator agreement. The recommended MT outputs are incorrectly recognized as TM outputs by one post-editor, which shows both the potential and the necessity for TM–MT integration.

This work can be extended in several ways. First of all, in this paper we concentrated on proprietary data and professional post-editors, according to the major paradigm in the localization industry. However, at the same time this limits the number of annotators we can hire, as well as the types of evaluations we can perform. We can obtain more comprehensive results by experimenting on open-domain data sets, and applying crowd-sourcing technologies such as Amazon Mechanical Turk[2] (Callison-Burch, 2009).

Secondly, during the evaluation we were able to collect a number of human judgements for training a new translation recommendation system. We plan to train a new recommendation model and to compare the difference with models trained on automatic metric scores, when we have collected more human-annotated data.

Finally, this experiment can also be extended by measuring the actual post-editing time instead of the judgement time, which can lead to a more precise approximation of reduced post-editing effort when using translation recommendation to integrate MT outputs into a TM system.

Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University. We thank the anonymous reviewers for their insightful comments.

References

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263 – 311.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *The 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP-2009)*, pages 286 – 295, Singapore.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. 20(1):37–46.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Joseph L. Fleiss. 1981. *Statistical methods for rates and proportions*. John Wiley & Sons.

Ana Guerberof. 2009. Productivity and quality in mt post-editing. In *MT Summit XII - Workshop: Beyond Translation Memories: New Tools for Translators MT*, Ottawa, Ontario, Canada.

Yifan He, Yanjun Ma, Josef van Genabith, and Andy Way. 2010. Bridging TM and SMT with translation recommendation. In *ACL 2010*, Uppsala, Sweden.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *The 1995 International Conference on Acoustics, Speech,*

[2] `https://www.mturk.com`

and Signal Processing (ICASSP-95), pages 181 – 184, Detroit, MI.

Philipp Koehn and Barry Haddow. 2009. Interactive assistance to human translators using statistical machine translation methods. In *The Twelfth Machine Translation Summit (MT-Summit XII)*, pages 73 – 80, Ottawa, Ontario, Canada.

Philipp. Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *The 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL/HLT-2003)*, pages 48 – 54, Edmonton, Alberta, Canada.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *The 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions (ACL-2007)*, pages 177–180, Prague, Czech Republic.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Hsuan-Tien Lin, Chih-Jen Lin, and Ruby C. Weng. 2007. A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*, 68(3):267–276.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *The 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, pages 295–302, Philadelphia, PA.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *The 41st Annual Meeting on Association for Computational Linguistics (ACL-2003)*, pages 160–167, Sapporo, Japan.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, pages 61–74.

Richard Sikes. 2007. Fuzzy matching in theory and practice. *Multilingual*, 18(6):39 – 43.

Michel Simard and Pierre Isabelle. 2009. Phrase-based machine translation in a computer-assisted translation environment. In *The Twelfth Machine Translation Summit (MT Summit XII)*, pages 120 – 127, Ottawa, Ontario, Canada.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *The 2006 Conference of the Association for Machine Translation in the Americas (AMTA-2006)*, pages 223–231, Cambridge, MA.

Lucia Specia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009a. Estimating the sentence-level quality of machine translation systems. In *The 13th Annual Conference of the European Association for Machine Translation (EAMT-2009)*, pages 28 – 35, Barcelona, Spain.

Lucia Specia, Craig Saunders, Marco Turchi, Zhuoran Wang, and John Shawe-Taylor. 2009b. Improving the confidence of machine translation quality estimates. In *The Twelfth Machine Translation Summit (MT Summit XII)*, pages 136 – 143, Ottawa, Ontario, Canada.

Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *The Seventh International Conference on Spoken Language Processing*, volume 2, pages 901–904, Denver, CO.

Midori Tatsumi. 2009. Correlation between automatic evaluation metric scores, post-editing speed, and some other factors. In *The Twelfth Machine Translation Summit (MT-Summit XII)*, pages 332 – 339, Ottawa, Ontario, Canada.