Multi-Engine and Multi-Alignment Based Automatic Post-Editing and its Impact on Translation Productivity

Santanu Pal[1], Sudip Kumar Naskar[2], Josef van Genabith[1,3]
[1]Saarland University, Saarbrücken, Germany
[2]Jadavpur University, Kolkata, India
[3]German Research Center for Artificial Intelligence (DFKI), Germany
{santanu.pal, josef.vangenabith}@uni-saarland.de
sudip.naskar@cse.jdvu.ac.in

Abstract

In this paper we combine two strands of machine translation (MT) research: automatic post-editing (APE) and multi-engine (system combination) MT. APE systems learn a target-language-side second stage MT system from the data produced by human corrected output of a first stage MT system, to improve the output of the first stage MT in what is essentially a *sequential* MT system combination architecture. At the same time, there is a rich research literature on *parallel* MT system combination where the same input is fed to multiple engines and the best output is selected or smaller sections of the outputs are combined to obtain improved translation output. In the paper we show that parallel system combination in the APE stage of a sequential MT-APE combination yields substantial translation improvements both measured in terms of automatic evaluation metrics as well as in terms of productivity improvements measured in a post-editing experiment. We also show that system combination on the level of APE alignments yields further improvements. Overall our APE system yields a statistically significant improvement of 5.9% relative BLEU over a strong baseline (English–Italian Google MT) and 21.76% productivity increase in a human post-editing experiment with professional translators.

1 Introduction

The term Post-Editing (PE) is defined as the correction performed by humans over the translation produced by an MT system (Veale and Way, 1997). It is often understood as the process of improving a translation provided by an MT system with the minimum amount of manual effort (TAUS Report, 2010). While MT is often not perfect, post-editing MT can yield productivity gains as post-editing MT output may require less effort compared to translating the same input manually from scratch. MT outputs are often post-edited by professional translators and the use of MT has become an important part of the translation workflow. A number of studies confirm that post-editing MT output can improve translators' performance in terms of productivity and it may positively impact on translation quality and consistency (Guerberof, 2009; Plitt and Masselot, 2010; Zampieri and Vela, 2014). The wide use of MT in modern translation workflows in the localization industry, in turn, has resulted in substantial quantities of PE data which can be used to develop APE systems.

APE (Knight and Chander, 1994) has been proposed as an automatic method for improving raw MT output, before performing actual human post-editing on it. The approach is based on collecting human corrected output of a first stage MT system and using this to train a system to correct errors produced by the MT system, possibly resulting in a productivity increase in the translation process. The advantage of APE relies on its capability to adapt to any black-box MT engine; i.e., upon availability of post-edited data, no incremental training or full re-training of the MT system is required to improve the overall translation quality of the first stage MT system that was involved in the post-edition data collection. APE assumes the availability of source texts (S_{ip}), corresponding MT output (T_{mt}) and the human post-edited (T_{pe}) version of T_{mt}, and APE systems can be modelled as an MT system between $S_{ip}\_T_{mt}$ (i.e., a joint representation of S_{ip} and T_{mt}) and T_{pe}. However, statistical APE (SAPE) systems can also be built

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2559–2570, Osaka, Japan, December 11-17 2016.

without the availability of S_{ip} using only sufficient amounts of parallel "target-side" T_{mt}–T_{pe} text within the statistical MT (SMT) framework.

Usually APE tasks focus on systematic errors made by MT systems - the most frequent ones being incorrect lexical choices, incorrect word ordering, incorrect insertion or deletion of a word. The system presented in this paper explores the use of system combination in APE. System combination in MT has been studied extensively (Matusov et al., 2006; Du et al., 2009; Pal et al., 2014), except in the context of APE. Here we use system combination architectures on three different levels: (i) sequential combination between first-stage system and APE, (ii) parallel combination of alignment systems at the level of the APE and (iii) parallel combination of APE MT systems (including the first stage MT system). More precisely, our approach makes use of a hybrid implementation of multiple alignment combination within phrase-based SAPE (PB-SAPE) and hierarchical PB-SAPE (HPB-SAPE) and a system combination framework (a multi-engine pipeline) – that combines the best translations from the enhanced PB-SAPE, HPB-SAPE and the raw MT output. The model takes T_{mt} as input and provides T_{pe} as output. As we also use the output of the original first stage MT system in some combination experiments, our set-up indirectly also uses S_{ip} information. System combination and hybrid word alignment strategies are commonly used in MT, however to the best of our knowledge the work presented in this paper is the first approach to APE that uses system combination and hybrid word alignment methods within the APE engine. System combination has been found to be a very useful technique in MT where translation hypotheses from multiple MT engines are available. Motivated by the success of system combination in MT, we applied system combination in APE. Similarly, the use of multiple word alignments has been shown to improve MT results (Pal et al., 2013). For our APE, alignments have to be produced on "monolingual" target-side data (T_{mt} and T_{pe}). A particular focus of our paper is to explore the performance of hybrid alignments based on combinations of statistical and edit-distance based aligners in this "monolingual" setting.

The remainder of the paper is organized as follows. Section 2 gives an overview of the related work. Section 3 describes the components of our SAPE system. Section 4 outlines the data and data preprocessing and the experimental setup. Section 5 presents the results of automatic and human evaluation, followed by conclusions and avenues for further research in Section 6.

2 Related Research

APE approaches cover a wide methodological range. Simard et al. (2007a) and Simard et al. (2007b) applied phrase-based SMT (PB-SMT) for post-editing that handles the repetitive nature of errors typically made by rule-based MT (RBMT) systems. The APE system was trained on the output of the rule-based system as the source language and reference human translations as the target language. This APE system was able to correct systematic errors produced by the RBMT system and reduce the post-editing effort. The approach achieved large improvements in performance not only over the baseline rule-based system but also over a similar PB-SMT used in a standalone mode. Denkowski (2015) proposed a method for real time integration of post-edited MT output into the translation model. He extracted a grammar for each input sentence and applied it to the model. Rosa et al. (2012) and Mareček et al. (2011) applied a rule-based approach to APE of English–Czech MT outputs on the morphological level. They used 20 hand-written rules based on the most frequent errors encountered in translation. The method efficiently corrects morpho-syntactic categories of a word such as number, case, gender, person as well as dependency labels. The inclusion of source-language information in APE is also useful to improve the APE performance (Béchara et al., 2011). To overcome data sparsity issues, Chatterjee et al. (2015) proposed a pipeline where the best language model and pruned phrase table are selected through task-specific dense features. Recently, a bidirectional recurrent neural network model of APE using T_{mt}–T_{pe} was proposed by Pal et al. (2016) which consists of an encoder that encodes the MT output into a fixed-length vector from which a decoder provides a post-edited (PE) translation. They reported statistically significant improvement over a strong first stage MT system baseline.

Various automatic or semi-automatic post-processing techniques to implement corrections of repetitive errors have been developed, although often the overall resulting MT output after APE still needs to be

post-edited by humans in order to produce publishable quality translation (Roturier, 2009; TAUS/CNGL Report, 2010). Even though MT and APE output often need human PE, it is often faster and cheaper to post-edit MT and APE output than to perform human translation from scratch.

System combination is a technology where multiple translation outputs from potentially very different MT systems are combined. System combination includes (i) hypothesis selection (Rosti et al., 2007a; Hildebrand and Vogel, 2010), (ii) confusion network based decoding (Matusov et al., 2006; Rosti et al., 2007b) and (iii) model combination (DeNero and Macherey, 2011). The confusion networks are built using backbone selection using either multiple hypotheses as backbones (Leusch and Ney, 2010) or a single backbone (Rosti et al., 2007b; Du et al., 2009) using TER (Snover et al., 2006) or BLEU (Papineni et al., 2002). These alignment metrics select the hypothesis that agrees most with the other hypotheses on average. System combination can improve translation quality significantly which motivated us to apply the strategy for the APE task.

Some of the research mentioned above studied the impacts of various factors and methods in APE on productivity gains. However, those studies were not conducted to observe PE effort in commercial environments. The focus of our study is twofold - to examine how existing word alignment techniques and a system combination framework can be intelligently used to improve monolingual APE, and whether the improvements in APE measured in terms of automatic evaluation metrics translate to measurable productivity gains in human post-editing in commercial translation workflows.

3 System Description

Our APE system consists of four basic components: (i) a target side mono-lingual hybrid word alignment model based on a number of alignment approaches, (ii) PB-SAPE, (iii) HPB-SAPE, and (iv) a system combination module (also including the first stage MT system). The SAPE systems are trained monolingually with Italian T_{mt} generated by Google Translate (GT) and the manually post-edited translations T_{pe}.

3.1 A Hybrid Word Alignment Model for Target Side APE

Previous research in MT demonstrates that a combination of information coming from multiple alignment models can improve translation quality. This can be achieved in different ways, e.g., by combining exactly two bidirectional alignments (Och, 2003; Koehn et al., 2003; DeNero and Macherey, 2011), combining an arbitrary number of alignments (Tu et al., 2012; Pal et al., 2013), by constructing weighted alignment matrices over 1-best alignments from multiple alignments generated by different models (Liu et al., 2009; Tu et al., 2011) etc. Below we apply an alignment combination model to APE.

Our hybrid word alignment method combines word alignments produced by three different statistical word alignment methods: (i) GIZA++ (Och and Ney, 2003) word alignment with grow-diag-final-and (GDFA) heuristic (Koehn, 2010), (ii) Berkeley word alignment (Liang et al., 2006), and (iii) Sym-Giza++ (Junczys-Dowmunt and Szał, 2012) word alignment, as well as two different edit distance based word aligners based on TER (Translation Edit Rate) (Snover et al., 2006) and METEOR (Lavie and Agarwal, 2007). We follow Pal et al. (2013) in combining word alignment tables, however, we additionally use 3-word consistent phrases to generate more alignment links (cf. Section 3.1.3). We integrate the word alignment obtained with this hybrid model into our PB-SAPE (Pal et al., 2015) and HPB-SAPE (Pal, 2015) models.

3.1.1 Statistical Word Alignment

GIZA++ is a statistical word alignment tool which implements IBM models 1–5, an HMM alignment model, as well as the IBM-6 model for covering many to many alignments. The Berkeley word aligner uses an extension of Cross Expectation Maximization and is jointly trained with HMM models. Sym-Giza++ is a modification of GIZA++ . It modifies the counting phase of each model of Giza++ allowing for updating the symmetrisized models between the iterations of the original training algorithm. Sym-Giza++ computes symmetric word alignment models with the capability of taking advantage of multi-processor systems.

3.1.2 Edit Distance-Based Word Alignment

We use two different kinds of edit distance based word aligners where alignments are based on edit distance style MT evaluation metrics – METEOR and TER.

METEOR Alignment: METEOR is an automatic MT evaluation metric which provides an alignment between a translation hypothesis H (i.e., MT output) and a reference translation R (in this case the PE translation). Given a pair of strings such as H and R to be compared the alignment is a mapping between words in H and R, which is built incrementally by the three sequences of word-mapping modules: (i) **Exact:** maps if the words are exactly the same (ii) **Porter stem:** maps if they are the same after stemming (iii) **WN synonymy:** maps if they are synonyms in WordNet. If multiple alignments exist, METEOR selects the alignment for which the word order in the two strings is most similar (i.e. having the fewest number of crossing alignment links). The final alignment is produced as the union of all stage alignments (e.g. Exact, Porter Stem and WN synonymy).

TER Alignment: TER is an edit distance based automatic MT evaluation metric that measures the ratio between the number of edit operations that are required to turn a H into R to the total number of words in R. The allowable edit operations include insertion (I), substitution (S), deletion (D) and phrase shifts (Sh). As a byproduct of finding the minimum edit distance, it produces an alignment between the hypothesis and the reference. In the monolingual SAPE task, we make use of TER alignment as a potential alignment between T_{mt} and T_{pe}. The TER alignment between a T_{pe} and T_{mt} is illustrated in Figure 1. Where, the vertical bar '|' represents a match and *I*, *D* and *S* represent the three post-editing operations – insertion, deletion and substitution, respectively.

TL_{mt}:	w_1	w_2	w_3	ϵ	w_4	w_5	w_6	w_7	w_8	w_9
	|	D	S	I	|	|	S	|	|	|
TL_{pe}:	$\bar{w}_1$	ϵ	$\bar{w}_2$	$\bar{w}_3$	$\bar{w}_8$	$\bar{w}_4$	$\bar{w}_5$	$\bar{w}_6$	$\bar{w}_7$	$\bar{w}_9$

Figure 1: TER based monolingual word alignment

3.1.3 Producing Additional Alignments for Edit-Distance Based Alignment

To generate additional alignment points between parallel sentence pairs, we perform phrase extraction (Koehn et al., 2003)[1] between T_{mt} and T_{pe}. We extract all phrase pairs, T_{mt} phrase (e) and T_{pe} phrase ($\bar{e}$), that are continuous and consistent with the edit distance based monolingual alignments. This phrase extraction process is performed individually for both TER and METEOR based alignments. A phrase pair $(e, \bar{e})$ is consistent with alignment a if Equation 1 is satisfied.

$$(\forall w_i \in e : (w_i, x) \in a \wedge x \in \bar{e}) \wedge \ (\forall \bar{w}_i \in \bar{e} : (y, \bar{w}_i) \in a \wedge y \in e) \tag{1}$$

Unaligned words in a phrase pair are aligned to all the phrase internal words in the other language. Figure 2 depicts the process of generating additional alignments where the solid links represent edit distance based alignments and the dashed links represent the newly established alignments. The newly established alignment points are added to the corresponding (i.e, TER or METEOR) alignment matrix.

3.2 Hybridization

Our method follows the following heuristic. We consider either of the alignments generated by GIZA++ with grow-diag-final-and heuristic (Koehn, 2010) (a_1), Berkeley aligner (a_2), or SymGiza++ (a_3) as the standard alignment since edit distance based TER (a_4) and METEOR (a_5) fail to align many words in the monolingual Italian MT–PE parallel sentences. From the five alignments a_1–a_5, we compute the alignment combination as follows.

[1]For this task, we use 3-words phrases.

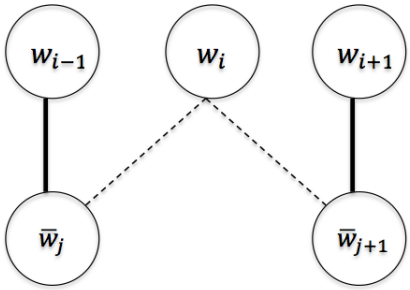

Figure 2: Producing additional alignments (w_i–$\bar{w}_j$, w_i–$\bar{w}_{j+1}$)

- **Step 1:** Choose a standard alignment[2] (S_a) from a_1, a_2 or a_3.
- **Step 2:** Produce a combined alignment $S_c = S_a \cup (a_2 \cap a_3)$, if a_1 is considered as S_a.
- **Step 3:** Delete all the alignment points $a_{ij} \in S_c$ such that $\exists a_{ik} \in a_4 \cup a_5$ where $j \neq k$.
- **Step 4:** Update S_c as $S_c = S_c \cup a_4 \cup a_5$.

3.3 System Combination for APE

Our system combination framework selects the best hypothesis translation from multiple hypotheses produced by different systems. In order to apply the system combination framework on the translations produced by our SAPE systems and the baseline MT system (Google Translate) we implemented the Minimum Bayes Risk (MBR) coupled with the Confusion Network (MBRCN) framework as described in (Du et al., 2009). The MBR decoder (Kumar and Byrne, 2004) selects for each sentence the best system output from the three outputs by minimizing BLEU (Papineni et al., 2002) loss. This output is known as the backbone. A confusion network (Matusov et al., 2006) is built from the backbone while the remaining hypotheses are aligned against the backbone using the edit-distance based alignment methods (cf. Section 3.1.2). The features used to score each arc in the confusion network (CN) are word posterior probability, target language model and length penalties. Minimum Error Rate Training (MERT) (Och, 2003) is applied to tune the CN weights. In our experiments, both APE hypotheses – PB-SAPE and HPB-SAPE, and the baseline Google Translate (GT) output are passed on to the system combination framework which produces the final system output (SC-APE).

4 Experiments

4.1 Data

The post-edition dataset for training the APE systems was developed in the MateCat[3] project. The data consist of 312K parallel sentences of Europarl and client data. The parallel data contains Italian translations (T_{mt}) produced by Google Translate from English as the source language and the corresponding post-edited Italian translations (T_{pe}) produced by professional translators. The parallel data were cleaned and processed by using a preprocessing module (see Section 4.2). After cleaning, we obtained a sentence-aligned MT–PE parallel corpus containing 213,795 sentence pairs. We randomly extracted 1,000 sentences each for the development set and test set from the parallel corpus and treated the remaining data (211,795) as the training set. The language model was built on the Italian Europarl corpus along with the PE side of the training set. The entire monolingual Italian corpus consists of 49,483,285 words.

4.2 Corpus Cleaning and Preprocessing

The MateCat corpus contains some non-Italian as well as non-English words and sentences. Therefore, we applied a language identifier (Shuyo, 2010) on both bilingual English–Italian MT output and MT

[2]Empirically best preforming aligner among the individual aligners (a_1, a_2 or a_3), is considered as S_a.
[3]https://www.matecat.com/

output–PE (Italian) parallel data. We discarded those sentence pairs from the bilingual training data which are considered as belonging to a different language or contain segment(s) in a different language. The same method was also applied to the monolingual Italian data. Next, the parallel corpus was further cleaned using the Gale-Church filtering method described in Tan and Pal (2014). We sorted the entire parallel training corpus based on sentence length and removed duplicates. We applied tokenization and punctuation normalization using the Moses scripts.

4.3 Experimental Settings

In our APE experiments we first integrated the hybrid word alignment model (cf. Section 3.1) into the SAPE engines modelled with PB-SMT (Koehn et al., 2003) and hierarchical PB-SMT (HPB-SMT) (Chiang, 2005). For building our statistical APE system, we used maximum phrase length of 7 and a 5-gram language model trained using KenLM (Heafield, 2011). Model parameters were tuned using MERT (Och, 2003) on the held-out development set.

5 Evaluation

During evaluation we take into consideration the output produced by all the three APE systems: PB-SAPE with hybrid word alignment, HPB-SAPE with hybrid word alignment and the system combination system (SC-APE) which also includes the output from the first stage system Google MT. As a baseline APE system we use a PB-SAPE system with GIZA++ alignment. The evaluation was carried out in two ways: (i) automatic evaluation and (ii) human evaluation of the 1,000 testset sentences automatically post-edited by our SAPE systems. Out of the 1,000 testset sentences, the output of the system combination based final post-editing system (SC-APE) were different from the raw Google Translate translation output for 198 sentences, i.e. only 19.8% of the GT translations are post-edited by the SC-APE system, the remaining sentences are not affected by APE. The entire testset is evaluated with automatic evaluation metrics while only the 198 sentences are subjected to human evaluation.

5.1 Automatic Evaluation

We evaluated the systems using three well known automatic MT evaluation metrics: BLEU, METEOR and TER. We also performed sentence level BLEU evaluation. Table 1 provides a comparison in terms of sentence level BLEU evaluation of the individual APE systems. Based on sentence level BLEU scores, the evaluation results presented in Table 1 show that 159 out of the 198 translations provided by the SC-APE are of better quality than the GT output. However, for the rest (39) of the translations, the GT output is of better quality than the APE output. This may be partly due to the fact that the human post-edited reference translations are biased towards GT output. However, manual analysis revealed that some of these 39 translations are indeed worse than the GT output. Overall, PB-SAPE, HPB-SAPE and SC-SAPE provide gains in terms of translation quality in 0.9%, 3.7% and 12% of the cases, respectively, as measured by S-BLEU. APE quality increases with the integration of the hybrid word alignment (HWA) model (cf. Section 3.2) into the different APE systems (cf. Table 2).

Systems	APE	GT	Tie	% Gain	% Loss
PB-SAPE	65	56	879	6.5%	5.6%
HPB-SAPE	91	54	855	9.1%	5.4%
SC-APE	159	39	802	15.9%	3.9%

Table 1: Automatic evaluation using Sentence-BLEU over 1,000 testset sentences.

Table 2 provides a comparison between the baseline PB-SAPE based on GIZA++ word alignment, PB-SAPE and HPB-SAPE based on hybrid word alignment (HWA), SC-APE and GT. The comparison is carried out in terms of BLEU, METEOR and TER scores. A general trend can be observed across all metrics. Baseline PB-SAPE system fail to improve over GT, while HWA based PB-SAPE, HPB-SAPE and SC-APE improve the translation quality over GT according to all metrics. Among the HWA based three APE systems, SC-APE performs best followed by HPB-SAPE and PB-SAPE in all metrics.

The SC-APE system provides 5.9%, 11% and 2.4% relative improvements over GT in BLEU, TER and METEOR, respectively, and all these improvements are statistically significant ($p < 0.01$). The HPB-SAPE system also provides promising improvements (4.2%, 7.3% and 1.2% in BLEU, TER and METEOR, respectively) over GT while PB-SAPE system yields in modest improvements.

Metric	PB-SAPE (Baseline)	PB-SAPE (HWA)	HPB-SAPE (HWA)	SC-APE	GT (First-Stage MT)
BLEU	59.90	62.70	63.87	**64.90**	61.26
TER	33.52	29.92	28.67	**27.52**	30.94
METEOR	69.54	73.31	73.63	**74.54**	72.73

Table 2: Automatic evaluation of the systems over 1,000 testset sentences.

5.2 Human Evaluation

The human evaluation process was carried out with 4 professional translators by introducing a polling system. The polling system provides every voter with three choices, two of which correspond to two different translation options for every source English segment. Translators act as voters and make a choice between the SC-APE output and the GT first-stage translation, based on whichever translation option looks better to them. Translators were also provided with a third option called *uncertain* (U), applicable whenever they are uncertain about which translation is better, i.e. when they deem both the GT and APE translations to be of equal quality (including equally unusable).

Table 3 shows the results of the polling scheme (human evaluation) of the raw GT output compared to the final automatic post-editing (SC-APE) output. The values in the table represent how many translations were chosen by each translator for individual systems. The polling based evaluation was carried out with 145 (of the 198) sentences. We discarded sentences containing less than six words either in source sentences or in the translations. We conducted the voting process serially to avoid any conflict between the translators. Table 3 shows that translators preferred APE output over the raw MT output. Translators did not have any knowledge about which translations are from which system as the two translation options were presented to them in random order. The winning APE system received on average 49.3% votes compared to 17% votes received by the GT system, while 33.7% votes were neutral as the translators were undecided for those sentences.

The SC-APE system received a total of 280 votes and it received votes from at least one translator for 105 unique segments, while GT received 112 total votes for 61 unique segments and 188 votes were received for 94 unique segments for the *uncertain* category. After detailed analysis we found that all 4 translators agreed on 27 APE translations, 6 GT translation and 9 neutral cases among the 145 sentences.

Translators	APE	GT	U
T1	91	22	32
T2	57	17	71
T3	72	37	36
T4	65	23	58
Average	71.5	24.7	49.2

Table 3: Outcome of polling with four expert translators for 145 sentences.

For the 145 sentences, we measured pairwise inter-annotator agreements between the translators by computing Cohen's κ coefficient (Cohen, 1960). Table 4 shows the inter-annotator agreements. The κ coefficients ranged from 0.141 (between T1 and T2) to 0.54 (between T2 and T4). The overall κ coefficient was 0.330. According to (Landis and Koch, 1977) this correlation coefficient can be interpreted as fair.

Cohen's κ	T1	T2	T3	T4
T1	-	0.141	0.424	0.398
T2	0.141	-	0.232	0.540
T3	0.424	0.232	-	0.248
T4	0.398	0.540	0.248	-

Table 4: Inter-annotator agreement between the translators.

5.3 Time and Productivity Gain Analysis

In order to investigate the performance of the APE system in terms of time and productivity gains, a completely new set of test data was distributed among the four translators. The new test data consists of real-life client segments. SC-APE and GT translations are presented separately to the translators within their daily usage interface (MateCat). Table 5 shows the statistics of how much time on average each individual translator took for the post-editing task. Table 5 also shows the average number of words (per minute, hour) post-edited by each translator. We calculated productivity gain by comparing column 2 (SC-APE) and 3 (GT) in Table 5. Table 5 shows that, SC-APE improves the productivity of the translators in general. Among the 4 translators, SAPE resulted in improved productivity for 3 translators (T1, T2 and T3), while for one translator (T4) it seems to result in productivity loss. If we look at the seconds/word, words/minute, and words/hour measures on the GT data for the 4 translators, it is easily noticeable that T1 is the most efficient post-editor, followed by T2, T4 and T3. However, when the translators work on the SAPE output, T2 is found to be the most productive while T4 is found to be the least productive. The productivity changes vary from 46.6% to -40%, which indicates that the utility of SAPE also varies from person to person. However, even taking into account the decrease in productivity of T4, average productivity increases 12.96% with SAPE. One thing to be noted here is that the productivity loss of T4 perhaps should not be considered for evaluation. We spoke to T4 after the evaluation and found that the translator was not solely concentrating on the post-editing job, switching among different jobs.

	SC-APE			**GT**				
	secs /word	**words /min**	**words /hour**	**secs /word**	**words /min**	**words /hour**	**Gain /hour**	**% Gain**
T1	2.81	21	1260	2.92	20	1200	60	5.0
T2	2.7	22	1320	3.88	15	900	420	46.6
T3	4.82	12	720	6.75	9	540	180	33.3
T4	9.80	6	360	5.84	10	600	-240	-40.0

Table 5: Post editing statistics over GT and SC-APE.

We also conducted a detailed evaluation of the post-editing carried out by the four translators. The results are reported in Table 6. Column 2 (fine grained evaluation score) in Table 6 shows the average of scores assigned to each translator by MateCat based on 5 criteria: tag issues (mismatches, white spaces), translation errors (mistranslation, additions/omissions), terminology and translation consistency, language quality (grammar, punctuation, spelling) and style (readability, consistent style and tone). MateCat also classifies each translator to one of the 4 performance levels[4] – excellent (3), acceptable (2), poor (1) and fail (0), for each of the above mentioned 5 criteria. Column 3 (weight based on quality) shows the sum of the scores indicating performance levels for the 5 criteria. By multiplying the values in column 2 and column 3, we arrive at the final assessment score assigned to each translator.

By weighting the percentage gain (cf. last column in Table 5) with the final assessment scores (cf. last column in Table 6), as in Equation 2, we obtain an average productivity increase of 21.76%. Even

[4] http://www.matecat.com/support/revising-projects/revising-translation-jobs/

	Fine grained Evaluation score (s_f)	Weight based on quality (w_q)	Final Assessment $fa = s_f \times w_q$
T1	4.46	7	31.22
T2	4.44	6	26.64
T3	1.33	2	2.66
T4	2.74	1	2.74

Table 6: Assessment of the post-editors based on their performance and quality.

considering the negative productivity of T4, this overall productivity gain is significant.

$$average\ productivity\ gain = \frac{\sum_{i=1}^{4} gain_i \times f_{ai}}{\sum_{i=1}^{4} f_{ai}} \quad (2)$$

6 Conclusions and Future Work

The use of a single statistical aligner in our PB-SMT based baseline APE fails to improve over raw Google MT output; instead it degrades the performance, as was also reported by (Béchara et al., 2011). This motivated us to use alignment combination models including both statistical and edit-distance based methods in our hybrid word alignment model for APE. By improving word alignment, the APE system automatically acquires better lexical associations and already the "hybrid alignment-based" PB-SAPE system shows improvements over the Google MT baseline. The reason for using a hierarchical phrase extraction model for APE is that it makes the model sensitive to syntactic structures. Moreover, HPB-SAPE captures global reordering by SCFG, helping to correct word order errors to some extent. Integration of our hybrid word alignment into the APE model resulted in both PB-SAPE (S_1) and HPB-SAPE (S_2) producing better translations than GT. System combination based APE (SC-APE) of S_1, S_2 and GT provided further statistically significant improvements over raw MT output. We performed statistical significance testing between GT, S_1, S_2 and SC-APE. S_1 provides statistically significant ($0.01 < p < 0.04$) improvements over GT across all metrics. Similarly S_2 yields statistically significant ($p < 0.01$) improvements over both GT and S_1 in all metrics. Our SC-APE system performs best and results in statistically significant ($p < 0.01$) improvements over all other systems across all metrics. In future, we will try bootstrapping strategies for further tuning the model and add more sophisticated features beyond the lexical level. The future study will also include a comparison of our system performance with Neural APE (Pal et al., 2016). We will also carry out experiments on different datasets including the WMT APE datasets.

Acknowledgments

We would like to thank all the anonymous reviewers for their feedback. We are also thankful to Translated SRL, Rome, Italy. They shared their data for the experiments and enabled the manual evaluation of our system. Santanu Pal is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement no 317471. Sudip Kumar Naskar is supported by Media Lab Asia, DeitY, Government of India, under the Young Faculty Research Fellowship of the Visvesvaraya PhD Scheme for Electronics & IT. Josef van Genabith is supported by funding from the European Unions Horizon 2020 research and innovation programme under grant agreement no 645452 (QT21).

References

Hanna Béchara, Yanjun Ma, and Josef van Genabith. 2011. Statistical Post-Editing for a Statistical MT System. In *Proceedings of MT summit XIII*, pages 308–315.

Rajen Chatterjee, Marion Weller, Matteo Negri, and Marco Turchi. 2015. Exploring the planet of the apes: a comparative study of state-of-the-art methods for mt automatic post-editing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 156–161.

David Chiang. 2005. A Hierarchical Phrase-based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

John DeNero and Klaus Macherey. 2011. Model-based Aligner Combination Using Dual Decomposition. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 420–429.

Michael Denkowski. 2015. *Machine Translation for Human Translators*. Ph.D. thesis, Carnegie Mellon University.

Jinhua Du, Yifan He, Sergio Penkale, and Andy Way. 2009. MATREX: The DCU MT System for WMT 2009. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 95–99.

Ana Guerberof. 2009. Productivity and quality in the post-editing of outputs from translation memories and machine translation. *Localisation Focus*, 7(1):133–140.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197.

Almut Silja Hildebrand and Stephan Vogel. 2010. Cmu system combination via hypothesis selection for wmt'10. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 307–310, Uppsala, Sweden, July. Association for Computational Linguistics.

Marcin Junczys-Dowmunt and Arkadiusz Szał. 2012. SyMGiza++: Symmetrized Word Alignment Models for Statistical Machine Translation. In *Proceedings of the 2011 International Conference on Security and Intelligent Information Systems*, pages 379–390.

Kevin Knight and Ishwar Chander. 1994. Automated Postediting of Documents. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (Vol. 1)*, pages 779–784.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA.

Shankar Kumar and William Byrne. 2004. Minimum Bayes Risk Decoding for Statistical Machine Translation. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 169–176.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–74.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231.

Gregor Leusch and Hermann Ney. 2010. The rwth system combination system for wmt 2010. In *ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pages 315–320.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by Agreement. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 104–111.

Yang Liu, Tian Xia, Xinyan Xiao, and Qun Liu. 2009. Weighted alignment matrices for statistical machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1017–1026, August.

David Mareček, Rudolf Rosa, Petra Galuščáková, and Ondřej Bojar. 2011. Two-step Translation with Grammatical Post-processing. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 426–432.

Evgeny Matusov, Nicola Ueffing, and Hermann Ney. 2006. Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment. In *Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–40.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 160–167.

Santanu Pal, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. 2013. A Hybrid Word Alignment Model for Phrase-Based Statistical Ma-chine Translation. *ACL 2013*, pages 94–101.

Santanu Pal, Ankit Srivastava, Sandipan Dandapat, Josef van Genabith, Qun Liu, and Andy Way. 2014. USAAR-DCU Hybrid Machine Translation System for ICON 2014. In *Proceedings of the 11th International Conference on Natural Language Processing*, Goa, India.

Santanu Pal, Mihaela Vela, Sudip Kumar Naskar, and Josef van Genabith. 2015. USAAR-SAPE: An English–Spanish Statistical Automatic Post-Editing System. In *Proceedings of WMT*, pages 216–221, Lisbon, Portugal.

Santanu Pal, Sudip Kumar Naskar, Mihaela Vela, and Josef van Genabith. 2016. A neural network based approach to automatic post-editing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 281–286, August.

Santanu Pal. 2015. Statistical Automatic Post Editing. In *The Proceedings of the EXPERT Scientific and Technological workshop*, pages 13–22.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318.

Mirko Plitt and François Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics*, 93:7–16.

Rudolf Rosa, David Mareček, and Ondřej Dušek. 2012. DEPFIX: A System for Automatic Correction of Czech MT Outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 362–368.

Antti-Veikko Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie J. Dorr. 2007a. Combining outputs from multiple machine translation systems. In *In Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*, pages 228–235.

Antti-Veikko Rosti, Spyros Matsoukas, and Richard Schwartz. 2007b. Improved word-level system combination for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 312–319.

Johann Roturier. 2009. Deploying Novel MT technology to Raise the Bar for Quality: a Review of Key Advantages and Challenges. In *Proceedings of the twelfth Machine Translation Summit*.

Nakatani Shuyo. 2010. Language Detection Library for Java.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007a. Statistical Phrase-based Post-editing. In *In Proceedings of NAACL*, pages 508–515.

Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. 2007b. Rule-based translation with statistical phrase-based post-editing. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 203–206.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231.

Liling Tan and Santanu Pal. 2014. Manawi: Using Multi-word Expressions and Named Entities to Improve Machine Translation. In *Proceedings of Ninth Workshop on Statistical Machine Translation*.

TAUS Report. 2010. Post editing in practice. Technical report, TAUS.

TAUS/CNGL Report. 2010. Maschine Translation Post-Editing Guidelines Published. Technical report, TAUS.

Zhaopeng Tu, Yang Liu, Qun Liu, and Shouxun Lin. 2011. Extracting Hierarchical Rules from a Weighted Alignment Matrix. In *In Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1294–1303.

Zhaopeng Tu, Yang Liu, Yifan He, Josef van Genabith, Qun Liu, and Shouxun Lin. 2012. Combining Multiple Alignments to Improve Machine Translation. In *The 24th International conference of computational linguistics (Coling 2012)*, pages 1249–1260.

Tony Veale and Andy Way. 1997. Gaijin: A Bootstrapping, Template-driven Approach to Example-based MT. In *Proceedings of the Recent Advances in Natural Language Processing*.

Marcos Zampieri and Mihaela Vela. 2014. Quantifying the Influence of MT Output in the Translators Performance: A Case Study in Technical Translation. In *Proceedings of the EACL Workshop on Humans and Computer-assisted Translation (HaCat)*, pages 93–98.