Improving Translations by Combining Fuzzy-Match Repair with Automatic Post-Editing

John E. Ortega
Universitat d'Alacant
E-03071, Alacant, Spain
jeo10@alu.ua.es

Marco Turchi
Fondazione Bruno Kessler
Trento, Italy
turchi@fbk.eu

Felipe Sánchez-Martínez
Universitat d'Alacant
E-03071, Alacant, Spain
fsanchez@dlsi.ua.es

Matteo Negri
Fondazione Bruno Kessler
Trento, Italy
negri@fbk.eu

Abstract

Two of the more predominant technologies that professional translators have at their disposal for improving productivity are machine translation (MT) and computer-aided translation (CAT) tools based on translation memories (TM). When translators use MT, they can use automatic post-editing (APE) systems to automate part of the post-editing work and get further productivity gains. When they use TM-based CAT tools, productivity may improve if they rely on fuzzy-match repair (FMR) methods. In this paper we combine FMR and APE: first a FMR proposal is produced from the translation unit proposed by the TM, then this proposal is further improved by an APE system specially tuned for this purpose. Experiments conducted on the translation of English texts into German show that, with the two combined technologies, the quality of the translations improves up to 23% compared to a pure MT system. The improvement over a pure FMR system is of 16%, showing the effectiveness of our joint solution.

1 Introduction

In recent times, research has shown that translators can be more productive when applying state-of-the-art post-editing techniques (Isabel, 2017). In many cases, the state-of-the-art techniques are applied to improve translation proposals from a translation memory (TM) or directly produced by a machine translation (MT) system. Post-editing techniques can be automated and seamlessly integrated into the typical translation pipeline for productivity gains. Two such techniques: fuzzy-match repair (FMR) (Ortega et al., 2016) and automatic post-editing (APE) (Chatterjee et al., 2017) have shown to be effective without the initial intervention of the translator by offering a *repaired* translation proposal from a TM in the case of FMR, and an improved MT output in the case of APE.

FMR is an automatic post-editing technique typically used with TM-based computer-aided translation (CAT) tools. In TM-based CAT, the translator is offered a translation proposal that comes from a translation unit (a pair of parallel segments) whose source segment is similar to the segment to be translated. When the source segment in the translation unit and the segment to be translated are not identical, which happens very often, the translation proposal needs to be post-edited in order to create the final translation. FMR aims to provide *repaired* translation hypotheses to reduce the post-editing effort of the original translation proposals by using another source of bilingual information such as an MT system. Some approaches to FMR, like the one by Koehn and Senellart (2010), heavily depend on the specific MT system type being used for repairing. Others, such as the one by Ortega et al. (2016) use an agnostic, black-box, MT system in such a way that the user would only choose from several repaired hypothesis proposals.

APE aims to correct the errors present in a machine-translated text before showing it to the translator or post-editor. As motivated by Parton et al. (2012), an APE system can help to improve MT output by exploiting information that is not available during translation, or by performing a deeper text analysis, and by adapting the output of

a general-purpose MT system to the lexicon/style requested in a specific application domain. In doing so, APE aims to provide professional translators with improved MT output quality to reduce (human) post-editing effort.

In this paper, we show that APE could be used to improve sentence-level proposals from FMR when FMR is used as a device to create new translations from a TM. As shown in Figure 1, FMR is first used to produce a repaired translation proposal and then APE is used as a tool to improve the quality of the proposal. We demonstrate that the combination of these two techniques can significantly boost translation quality. It outperforms both a competitive neural MT system and FMR alone, and its performance reaches nearly that of methods relying on the reference (i.e. oracle) translations.

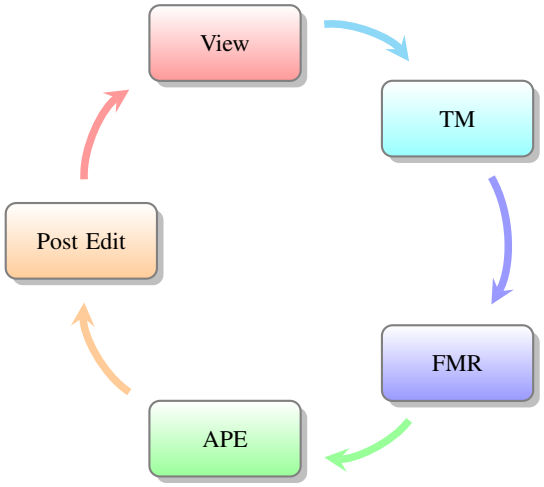

Figure 1: Seamless addition of fuzzy-match repair (FMR) and automatic post-editing (APE) in a traditional computer-aided translation (CAT) pipeline. The post editor is presented with several hypotheses created from a translation memory (TM) proposal through fuzzy-match repair (FMR) and automatic post-editing (APE).

Our work provides an in-depth analysis of which technique would work best under "typical" translation scenarios by testing several combinations of the two post-editing techniques. Our analysis includes various checkpoints of evaluation including industry standards and human-level reviews. In order to better describe our process, we organize the paper as follows. First, in Section 2 we review the relevant work where both technologies (FMR and APE) have been used. Second, in Section 3, we dig deeper into the motivation and methodology of our work and show how the two technologies could be "glued" together to form a new system that is added in a modular way to a traditional CAT pipeline. Third, in Section 4 we describe our experimental settings in detail. Fourth, we present our results in Section 5. We use BLEU and TER as metrics to evaluate the quality of our translations. We also perform error analysis and human reviews. Then, we measure the systems quantitatively using a word-measurement like word-error rate to show performance. Finally, in Section 6 we give some conclusions and plan on doing in the future.

2 Related work

In this section we describe approaches related to both FMR and APE. It is worth noting that, to the best of our knowledge, FMR and APE have not previously been combined together.

2.1 Fuzzy-match repair

FMR aims to reduce the post-editing effort of translation proposals retrieved from a TM. To do so FMR techniques rely on a source of bilingual information, usually MT, to automatically *repair* a translation proposal by modifying those parts of the proposal that otherwise should be post-edited by the translator. The idea of FMR points back to papers by Kranias and Samiotou (2004) and Hewavitharana et al. (2005) whose approaches were based on the location of anchor points via alignment of words and relied heavily on the inner workings of the MT system they used. Improvements over time led way to advances that used phrase-based MT (Simard and Isabelle, 2009; Koehn and Senellart, 2010). Work has gradually advanced and various FMR methods have been proposed that share one common theme: locating and repairing sub-segments in the translation proposal. Later works (Dandapat et al., 2011; Ortega et al., 2016), on the other hand, can use *any* MT system as a black-box.

Knowles et al. (2018) recently performed a comparison of the nature of MT systems for their use in FMR. In particular, they contrast the quality of FMR output using neural MT and phrase-based MT. Most importantly, they show that neural MT may not be appropriate if it is not trained on in-domain data. Other novel works, like the work by Bulté et al. (2018), include FMR as a primary part of a system integrating MT and TM. Lastly, Ortega et al. (2018) have found a statistical way to select the best MT system to use in black-box FMR.

2.2 Automatic post-editing

Automatic post-editing is the task of correcting recurring errors from an MT system by learning from human corrections. Starting from the seminal work by (Simard et al., 2007), the problem has been tackled as a "monolingual translation" task in which the MT output must be translated into an improved text in the target language. Under this definition, the "parallel data" used for training an APE system consist of triplets of the form (source, target, post-edited target) rather than the (source, target) pairs normally used in MT. Following the translation-based approach, initial solutions relied on the phrase-based paradigm (Simard et al., 2007; Dugast et al., 2007; Terumasa, 2007; Pilevar, 2011; Béchara et al., 2011; Chatterjee et al., 2015; Chatterjee et al., 2016). However, in the past couple of years, top results have been achieved by neural architectures (Pal et al., 2016; Junczys-Dowmunt and Grundkiewicz, 2016; Chatterjee et al., 2017; Junczys-Dowmunt and Grundkiewicz, 2017; Tebbifakhr et al., 2018).

Recent advancements made by participants in the APE shared task organized within the Conference on Machine Translation (WMT) have shown the capability of APE systems to significantly improve the performance of a black-box MT system gaining up to seven BLEU points (Bojar et al., 2017; Chatterjee et al., 2018a).

The neural approaches proposed share common traits such as using multi encoders (one for the source and one for the MT segments) and leveraging artificial data (round-trip translations) to maximize results. The APE system used in this paper proposes a novel approach extending the original technology implemented by the best performing system at the WMT 2016 APE shared task (Chatterjee et al., 2017).

2.3 Combination of approaches

We briefly describe a few combinations of approaches and systems that are usually used in different scenarios, such as FMR and APE, and that could be considered novel and related to our work. The first, and probably the most relevant work, is based on MT quality estimation (QE) and APE. Chatterjee et al. (2018b) combine MT QE and APE in three different ways: one in which sentence-level MT QE is used to activate an APE system, a second one in which word-level MT QE is used to guide the APE system, and a third one that uses MT QE to choose between the original MT output and its post-edited version. Additionally, Tan et al. (2017) attempts to correct a common problem in APE known as "overcorrection" (i.e. systems' tendency to completely re-translate the MT output, also rephrasing parts that are already correct). They do this by specifying two models (called neural post-editing models). Then, they use MT and QE to help select one of the models for the translation. This by no means is related to fuzzy-match repair; however, the idea of combining several systems around APE is similar to what we are doing.

Hokamp (2017) includes word-level MT QE features as additional inputs to an APE system and trains several neural models using different input representations, but sharing the same output space. These models are finally ensembled together and tuned for APE and MT QE.

3 TM repairing through FMR and APE

Our system is a two-step process that can be added to any TM-based CAT tool that has access to a source of bilingual information (SBI), such as a black-box MT system. The first step of our process is to use the translation unit whose source segment is most similar to the segment to be translated as input to FMR that, in turn, uses the SBI for repairing and proposing new translation hypotheses not present in the TM. These proposals could then be treated as input to a second APE step that is used to output the best final possible hypotheses. In this section, we first describe more formally how FMR and APE are used. Then, we provide an example (Table 1) in the last sub-section that illustrates how APE can be used to improve an FMR translation proposal.

3.1 Fuzzy-match repair

The FMR method devised by Ortega et al. (2016) can generate a set of fuzzy-match repair hypotheses from a translation unit (s, t) and the segment to be translated s' by using any available SBI. For our experiments, we use MT[1] as a black-box SBI.

Their method first identifies mismatched words between s and s', that is, the words they do not have in common. This is done by using the alignment between the words in s and s' obtained as a by-product of the computation of the word-based edit distance (Levenshtein, 1966) between s and

[1]Other SBIs that could be used are sub-segment translation memories, bilingual dictionaries or phrase tables.

s': mismatched words are left unaligned. SBIs are then used to translate into the target language sub-segment pairs of s and s' containing mismatched words. The sub-segments pairs to be translated are obtained by using the phrase-pair extraction algorithm used in phrase-based statistical MT to obtain bilingual phrase pairs (Koehn, 2010, section 5.2.3). The translations obtained for the sub-segments of s are used to identify the sub-segment in t that needs to be modified, and the translation of the sub-segments of s' to identify the way they should be modified. In this way, a set of *patching operators* is built. Each patching operator consists of a sub-segment σ of s, a sub-segment σ' of s' aligned with σ, a sub-segment τ of t to be repaired, and a sub-segment τ', the translation of σ', to be used for repairing. By combining these patching operators, a set of fuzzy-match repaired hypothesis is generated. For a detailed description of their method, we refer the reader to the work by Ortega et al. (2016).

3.2 Automatic post-editing

The APE system used in this paper is a re-implementation of the multi-source attention-based encoder-decoder system (Chatterjee et al., 2017) that achieved the best performance in the automatic evaluation at the APE shared task at WMT 2016.[2] This system uses two different encoders to independently process the source and the MT segments. Each encoder consists of a bi-directional GRU and has its own attention layer that is used to compute the weighted context. To obtain a single context, the two context vectors are combined via a feed-forward network. The obtained context is used to compute the classical attention model (Bahdanau et al., 2015). To regularize the multi-source network and to avoid over-fitting, a shared dropout is applied to the hidden state of both encoders and to the merged context. This architecture has shown to be particularly effective in the APE task, and its multi-source structure makes it particularly suitable for the FMR post-editing task.

3.3 FMR with APE

The integration of FMR and APE does not require that the two ideas share any code behind the scenes; so, both can be seen as black box mechanisms for improving translation proposals from the TM. For this paper, FMR first creates several new proposals based on the original TM proposals. Then, APE uses those proposals as the base to produce even better proposals.

[2] http://www.statmt.org/wmt16/

Source:	article 18 , paragraph 1 , of the co2 act
TM:	article **45** , paragraph 1 , of the co2 **ordinance**
FMR:	artikel 18 absatz 1 der co2-verordnung
APE:	artikel 18 absatz 1 des co2-gesetzes
Ref:	artikel 18 , absatz 1 des co2-gesetzes

Table 1: An example of how fuzzy-match repair (FMR) and automatic post-editing (APE) could work together to improve a translation memory (TM) proposal.

Table 1 shows an example of how a source sentence from our TM is modified first by FMR and then by APE. First, FMR repairs the TM proposal by replacing two words (*45* and *ordinance*); notice that FMR incorrectly translates *co2 act* as *co2-verordnung*. APE then takes the FMR proposal and produces an improved translation, *co2-gesetzes*, which is closer to the reference translation. The final result is a more adequate translation that needs fewer post-edits by the final user.

4 Experimental Settings

We experiment with various combinations of FMR and APE using a phrase-based MT system as a SBI for FMR. In addition, we use APE on the output of two MT systems, a phrase-based MT system and a neural MT system, as a point of comparison. This section goes over the details of the data and systems we used. One of our goals in this paper is to show that by using freely-available data found on the Internet, which is the case for small businesses that do not have in-house data and cannot afford more expensive data sets, our system achieves good results despite results from previous work (Knowles et al., 2018; Chatterjee et al., 2018b) that have shown that training MT systems on in-domain data, especially in the case of a neural MT system, can be advantageous.

4.1 Data

Our entire dataset is based on 4,000 randomly selected sentences from the DGT translation memory (DGT-TM-release 2018).[3] The TM is available in several languages containing many translation units.[4] In our evaluation, we use the English–German (EN–DE) TM extracted with the formal

[3] ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory

[4] For some statistics about this TM, please visit wt-public.emm4u.eu/Resources/DGT-TM_

DGT extraction methodology mentioned on their website.

FMR is used to generate repaired translation hypotheses for these 4,000 sentences by using the whole EN–DE DGT TM to look for translation units to repair; it is worth noting that the whole DGT TM is not used in any way by the APE system. FMR hypotheses are generated for each of the 4,000 segments by looking in the whole DGT TM for the translation unit (s, t) whose source segment s is the most similar to the segment to be translated s'. The similarity between s' and s is computed as the *fuzzy match score*, which in turn is based on the word-based edit distance (Levenshtein, 1966) between s and s'. If a translation unit with a fuzzy match score above 60%[5] is found, it is used for FMR; otherwise, the Moses (Koehn et al., 2007) MT system is used to translate s'.

Of the 4,000 sentences selected at random from the DGT TM, 2,500 are randomly selected and used to fine-tune the APE system (see Section 4.3), 500 are used for development, and 1,000 for testing. Altogether, about 350 sentences are not successfully repaired by FMR; in those cases, we used the output of Moses.

4.2 Machine translation systems

We use the phrase-based statistical MT system Moses (Koehn et al., 2007) as a SBI for FMR; it has shown to perform well in previous experiments and in the black-box setting (Knowles et al., 2018). As a term of comparison we use Moses and the neural MT system Nematus (Sennrich et al., 2016) as baselines; we leave for future work the inclusion of a neural MT system as a SBI for FMR. It is worth noting that the phrase-based MT system performed better on the APE module than the neural MT system (see Table 2).

With Moses we use pre-trained models downloaded from `www.statmt.org/moses/RELEASE-3.0/models/`. By using pre-trained models, we try to replicate what most users in a corporate setting would choose, at least as a first iteration, in absence of advanced knowledge to build the MT models by their own.

Nematus is trained on a collection of datasets belonging to different domains. This is done to resemble a typical industrial scenario where a translation system is trained on a large collection of data that may or may not match the test domain. In particular, we use domain-specific parallel corpora from the European Central Bank, Gnome, JRC-Acquis, KDE4, OpenOffice, PHP and Ubuntu,[6] and generic training sets obtained from the CommonCrawl dataset[7] and Europarl.[8] The Europarl corpus can be considered an in-domain dataset because it belongs to the same domain of the DGT TM collection.

To train Nematus, the training corpus is first processed using byte pair encoding (BPE) (Sennrich et al., 2016), so that the less frequent words are segmented into their sub-word units, resulting in vocabularies of maximum size of 90k entries, or 90k BPE operations. The size of word embeddings and hidden layers is set, respectively, to 500 and 1024. Source and target dropout are both set to 10%, whereas, encoder and decoder hidden states and embedding dropout is set to 20%. The learning rate is set to 0.001. The cost is computed on mini-batches of 100 sentence pairs with maximum length of 50 tokens, extracted from the randomly shuffled data after each epoch. The models are optimized using Adagrad (Duchi et al., 2011) and every 10,000 mini-batches they are evaluated with BLEU on the 500-sentence-pairs development set.

4.3 APE settings

The APE system is trained on the eSCAPE corpus (Negri et al., 2018), a collection of ∼7M triplets (source, MT output and reference), where the MT outputs have been created by a phrase-based MT system. It consists of datasets belonging to different domains and it is filtered by removing duplicates and too short (3 words) or too long (60 words) segments.

To adapt the generic APE system to the FMR task, the model is fine-tuned (Luong and Manning, 2015) on 2,500 triplets (see Section 4.1), where the source input is paired with the repaired translation proposal produced by FMR.

Similar to the neural MT system, the APE system is trained on sub-word units by using BPE. The APE vocabulary is created by selecting 50k most frequent sub-words. Word embedding and GRU hidden state size is set to 1024. Network parameters are optimized with Adagrad with a learning rate of 0.01. Source and target dropout is set to

`Statistics.pdf`

[5] We use 60% fuzzy-match as a starting point threshold; in future work, we plan on trying with higher thresholds.

[6] All available at `opus.lingfil.uu.se`.

[7] `www.statmt.org/wmt13/training-parallel-commoncrawl.tgz`

[8] `www.statmt.org/europarl/`

10%, whereas, encoder and decoder hidden states, weighted source context, and embedding dropout is set to 20%. After each epoch, the training data is shuffled and the batches are created after sorting 2,000 samples in order to speed-up the training. The batch size is set to 100 samples, with a maximum sentence length of 60 sub-words. The fine-tuning step is performed using the same parameters of the generic training.

4.4 Combined FMR and APE settings

Our FMR approach is identical to the FMR approach presented by Ortega et al. (2016). The only things that change are the MT system used as SBI, the language pair and the TM used. The output produced for experimentation by FMR is a list of translated segments that serve as input to the APE system. In particular, we experiment with two main FMR outputs for APE integration:

- an **oracle** experiment that chooses the best possible repaired translation hypothesis for each segment s' by computing the word-based edit distance between the repaired translation and the reference translation;

- a **randomized** experiment that, for each segment s', choses at random a repaired translation from the whole set of repaired translation hypotheses. On average, there are nearly 5 hypotheses per source segment s'. We use a random selection method because of its simplicity and because the chance of choosing the best hypothesis is around 20%.

4.5 Evaluation setting

For evaluating the combination of FMR and APE, we use two major metrics: BLEU (Papineni et al., 2002) and translation edit rate (TER) (Snover et al., 2006). We report on BLEU because it is a centerpiece of the development of MT systems, and on TER because it is the primary evaluation metric at the WMT APE shared task.

In addition to automatic evaluation metrics, we introduce a human evaluator: a native German speaker. This evaluator is not a translator; yet, does have a background in natural language processing and evaluation.[9] We report the evaluator's overall evaluation on the best performing systems in our results and offer it as an extra evaluation metric of performance. The hope is to better understand the target language and how well the various systems perform under a native eye.

We provided a random set of 1,000 samples to the evaluator, where each sample is made of a sentence pair and its translations provided by each system presented in Table 3. Each sentence pair is rated by assigning quality scores on a 5-point scale (1 being the worst and 5 the best). The evaluator was told to rate the quality of translations and, thus, was given the final translation from the four systems but not the original human reference translation. Additionally, the evaluator was asked to provide an explanation of why each system's translation did not seem correct. Correctness was determined as a system's translation being exactly what was expected for the source sentence (a 5-star rating) or not at all (a 1-star rating).

5 Results

In this section we present results broken down into two different sub-sections to highlight the performance of the final combination system from the 1) system level and 2) human perspective. In Section 5.1 we report two major MT metrics: BLEU and TER. Then, in Section 5.2, the evaluator's feedback is taken into account while analyzing specific text anomalies that were found in the evaluation.

5.1 Metric-based analysis

Table 2 shows results that compare the use of MT, FMR, and APE for translation. They contain two main FMR configurations: **FMR Rand** – selecting a translation hypothesis at random, and **FMR Oracle** – using the hypothesis from FMR that is the nearest to the reference translation in terms of word-error rate. We also provide three variants obtained by combining FMR with APE (**FMR-APE**; see Section 4.4). The first three rows of Table 2 represent baseline experiments without the use of FMR or APE. We consider them as our baseline experiments because they are: the output of the phases-based MT system Moses (**PBMT**), the neural MT system Nematus (**NMT**), and the translation proposal as found in the translation memory (**TM**). APE is then measured alone using the two MT systems (PBMT and NMT) in the two rows **Phrase-based MT-APE** and **NMT-APE**. FMR alone is evaluated after that in the **FMR RAND** and **FMR Oracle**

[9] For economic and timing reasons, we only present evaluation from a single evaluator.

System	BLEU	TER
PBMT	39.62	49.74
NMT	51.54	36.75
TM	64.95	25.42
Phrase-based MT-APE	60.02	31.60
NMT-APE	56.58	33.77
FMR Rand	58.38	32.17
FMR Oracle	68.36	23.03
FMR-APE Rand	66.56	26.20
FMR-APE Oracle	80.54	15.60
FMR-APE Oracle-Rand	**74.44**	**20.26**

Table 2: Performance of three baseline approaches (use of a phrase-based MT system, use of a neural MT system, and use of the TM proposal without repairing), of the use of APE to better the MT outputs, the use of FMR alone when the translation hypothesis is selected at random or using an oracle, and of different combinations of FMR and APE.

rows. Then, the combination of FMR and APE with a random FMR hypothesis choice and an oracle (**FMR-APE Rand** and **FMR-APE Oracle**) is presented. Lastly, we present **FMR-APE Oracle-Rand**, which is our best approximation of FMR with APE that uses the randomly chosen hypothesis from FMR for each source segment as additional training data to the APE system.

The TM baseline approach performs the best when compared to the two MT systems (+∼25 BLEU points over the phrase-based MT and +∼13 over the neural MT). We attribute the performance of the TM approach to the fact that the DGT-TM is highly repetitive: it is quite likely that a match with a high fuzzy match score is found. The TM matches account for more than 70% of the 1,000 test segments; that is, for 70% of the segments there is a translation unit for which the fuzzy match score is above 60%. The TM baseline does quite well when matched; and, when it is not matched, Moses is used to translate the entire sentence.

FMR Rand is significantly below the TM approach, showing that there is a need for a better strategy to choose the best FMR repaired hypothesis in absence of a reference translation to propose to a post-editor. Selecting from hypotheses at random in FMR can generate low-quality segments that could reduce a post-editor's trust in the method. With the oracle selection (FMR Oracle), we notice a significant boost in performance (+4 BLEU points over the TM and +10 over the FMR Rand method). However, the oracle solution should only be considered as an upper bound for optimum FMR hypothesis selection purposes. We leave a better selection method for FMR based on quality estimation for future work.

When combining FMR with APE, in both cases (FMR-APE Rand and Oracle) and by a large margin (+8 BLEU points for Rand and +12 for Oracle), APE improves translation quality with respect to FMR alone. The APE gain allows the FMR Rand method to also outperform the TM approach. At a closer look, APE seems to have a larger effect on the FMR Oracle than on FMR Rand. We believe that the random selection of hypotheses produces segments with few common characteristics that make it harder for APE to learn a strict correction pattern. For validation, we use the FMR-APE Oracle model as a training mechanism for APE because it contains hypotheses chosen by looking at the reference (FMR-APE Oracle-Rand in Table 2).[10] Results when using the FMR-APE Oracle as training for the APE model are the best and outperform both TM and FMR-APE Rand (+10 and +8 BLEU points). We consider this to be the best adaptation of FMR.

APE gains can be classified into two main categories: (1) addition of missing parts and (2) lexical substitution. In the former, since APE accepts the source and the MT sentences, APE inserts parts that are not present in the FMR output. In one example, the source sentence *"30 October 2015"* is translated by the FMR as *"30 2015"*, discarding the word *"October"* that is re-inserted by the APE system, thus matching the reference sentence *"30 Oktober 2015"*. The latter category (lexical substitution) is mainly related to the identification of the correct word and it is very important when dealing with one or more TMs, where two suggestions can only differ by one word. In another example, the source sentence *"Regulation 2015 / 7"* is translated by the FMR as *"Verordnung 2015 / 8"*, introducing a wrong number for the month. Leveraging the source, the APE is able to set the correct value matching the reference *"Verordnung 2015 / 7"*.

We report also on word-error rate (WER) in Figure 2 to get a better idea of how many words were actually modified by each system. Interestingly, the WER by most of the systems does not beat the TM score. We believe that this is due to the fact that the TM score is actually a mix of the TM and the phrase-based MT system; recall that Moses is

[10] Note that this strategy can be used in production because the training data relies on parallel data where the reference/oracle translation is available.

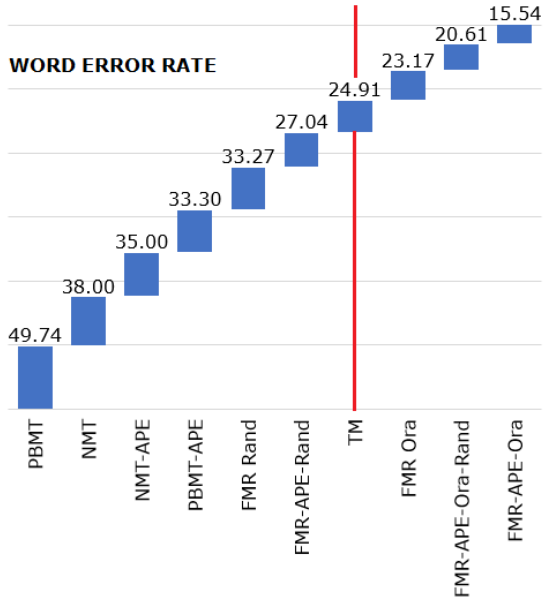

Figure 2: Word error rate (WER) for all of the systems. The best scoring system according to Table 2 (the FMR-APE Oracle-Rand system) also performs best according to the WER score of 20.61 on the top right.

Best System	Human Rating
TM	2.84
Phrase-based MT-APE	2.82
FMR Oracle	2.90
FMR-APE Oracle-Rand	3.67

Table 3: Average human evaluation for the best system combining FMR and APE. Translations were rated using a 5-point scale, 1 being the worst and 5 the best.

used when a good-enough translation unit is not found. Nonetheless, the best scoring systems are the FMR-APE combination systems.

5.2 Human-based analysis

The three measurements (BLEU, TER, and WER) show how well our best system performs and would probably be enough to show that it is worthwhile to combine FMR with APE. However, we passed the translations from our best performing systems to a native German evaluator (non-professional) well-versed in machine translation and natural language processing. Table 3 shows a quick overview of how the best systems perform: the human evaluation score is in line with the automatic metrics reported above.

We also asked the human evaluator to provide general comments on each of the best-performing systems. We did this to get a better idea of the types of errors each system made. Below is an overview of what the evaluator found.

TM. The most common error, accounting for nearly 30% of the incorrect cases from the TM, was "missing" or "wrong" data which describes typical information in the parliamentary texts like an article changing from 33 to 45. This is one of the reasons that a translator would like to use a TM because the translator would typically only have to change the numbers in those situations. There are also a few comments such as "wrong" part-of-speech, e.g. an adjective or noun being wrong.

Phrase-based MT-APE. Unlike the TM, we see some common phrase-based MT mistakes such as "noun cases wrong" that account for more than 15% of the total incorrect words. Also, since Moses marks untranslatable words as "UNK", we find that the evaluator noticed those anomalies made up 20% of the word-based issues. In addition to the normal mistakes, the evaluator noticed that on the order of 35% of the translations just "did not make sense", even more than the TM. That could be coupled with another finding, "repetition", to form what seems to be somewhat common in phrase-based MT-backed APE systems.

FMR Oracle. The best FMR was not immune to issues either. This could be due to the MT systems used. Many of the errors were similar to the Phrase-based MT-APE system; however, other errors were reported such as "punctuation is weird" and "important" words are missing. However, in more cases than others, it seems that the "FMR Oracle" system gets the underlying meaning correct.

FMR-APE Oracle-Rand. This system performed the best in all cases. We consider this to be the most important finding of this paper. While there were comments concerning UNK symbols (typical of the phrase-based MT translations), we saw some issues of morphology such as problems with inflection. For the most part, the evaluator made few comments because the translations were easier to understand than all other systems.

6 Conclusion

In this paper, we proposed a two-step process able to generate improved translations. The approach relies on the combination of two techniques: fuzzy match repair (FMR) and automatic post-editing (APE). Given a translation unit and the segment to be translated, the FMR module creates a set of

fuzzy-match repair hypotheses. The selected hypothesis is then fed as input to the APE system that fixes its errors. When compared against neural MT, a TM-based approach and FMR alone, the combined solution outperforms all these methods indicating the effectiveness of the proposed technique. We measure performance using common, industry-wide MT performance metrics: BLEU and TER. We also show how WERs for our experiments nearly correlate with the BLEU and TER scores. In addition to BLEU, TER, and WER, we provide a human rating from a native German speaker as insight into how the best-performing systems fair to the average reader (not necessarily a translator). By combining FMR and APE, we provide easy, seamless access to FMR and APE for translators and post-editors.

We believe that the combination of two orthogonal technologies like FMR and APE could improve most stand-alone post-editing systems. We have been able to get decent gains by seamlessly juxtaposing two post-editing techniques in a straightforward way. Clearly, other system combinations (including using APE before FMR or even with the TM) should be tried along with the introduction of other language pairs as is done in the original FMR work (Ortega et al., 2016).

Along this direction, in future we plan on going the next step by combining yet another system with FMR and APE: quality estimation. One can easily imagine how quality estimation could be used both as a precursor and a post-validator for FMR and APE. Lastly, we will also use both MT systems as SBIs for FMR to increase the coverage and the chances to build successful patching operators, and a quality-estimation inspired approach to select the best hypothesis among the set of hypotheses produced by the FMR method used.

Acknowledgements

We thank Katharina Kann for providing native German translation ratings for several MT systems during evaluation. John E. Ortega's work was partially supported by the Universitat d'Alacant. Felipe Sánchez-Martínez's work was founded by the Spanish Government through the EFFORTUNE project (project number TIN2015-69632-R).

References

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Béchara, Hanna, Yanjun Ma, and Josef van Genabith. 2011. Statistical post-editing for a statistical mt system. In *MT Summit*, volume 13, pages 308–315.

Bojar, O., R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, S. Huang, M. Huck, P. Koehn, Q. Liu, V. Logacheva, C. Monz, M. Negri, M. Post, R. Rubino, L. Specia, and M. Turchi. 2017. Findings of the 2017 conference on machine translation (wmt17). In *WMT17*, pages 169–214.

Bulté, Bram, Tom Vanallemeersch, and Vincent Vandeghinste. 2018. M3tra: integrating tm and mt for professional translators.

Chatterjee, Rajen, 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)*, volume 2, pages 156–161.

Chatterjee, Rajen, José G. C. de Souza, Matteo Negri, and Marco Turchi. 2016. The FBK Participation in the WMT 2016 Automatic Post-editing Shared Task. In *Proceedings of the First Conference on Machine Translation*, pages 745–750, Berlin, Germany, August. Association for Computational Linguistics.

Chatterjee, Rajen, M. Amin Farajian, Matteo Negri, Marco Turchi, Ankit Srivastava, and Santanu Pal. 2017. Multi-source Neural Automatic Post-Editing: FBK's participation in the WMT 2017 APE shared task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 630–638, Copenhagen, Denmark, September. Association for Computational Linguistics.

Chatterjee, Rajen, Matteo Negri, Raphael Rubino, and Marco Turchi. 2018a. Findings of the WMT 2018 Shared Task on Automatic Post-Editing. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 723–738, Belgium, Brussels, October. Association for Computational Linguistics.

Chatterjee, Rajen, Matteo Negri, Marco Turchi, Frédéric Blain, and Lucia Specia. 2018b. Combining quality estimation and automatic post-editing to enhance machine translation output. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, volume 1, pages 26–38.

Dandapat, Sandipan, Sara Morrissey, Andy Way, and Mikel L. Forcada. 2011. Using example-based MT to support statistical MT when translating homogeneous data in a resource-poor setting. In *Proceedings of the 15th conference of the European Association for Machine Translation*, pages 201–208. Leuven, Belgium.

Duchi, John, Elad Hazan, and Yoram Singer. 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*.

Dugast, L., J. Senellart, and P. Koehn. 2007. Statistical post-editing on systran's rule-based translation system. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 220–223.

Hewavitharana, Sanjika, Stephan Vogel, and Alex Waibel. 2005. Augmenting a statistical translation system with a translation memory. In *Proceedings of the 10th conference of the EAMT on 'Practical applications of machine translation'*, pages 126–132, Carnegie Mellon University, Pittsburgh, USA.

Hokamp, C. 2017. Ensembling factored neural machine translation models for automatic post-editing and quality estimation. In *WMT17*, pages 647–654.

Isabel, Lacruz, 2017. *Cognitive Effort in Translation, Editing, and Post-editing*, chapter 21, pages 386–401. Wiley-Blackwell.

Junczys-Dowmunt, M. and R. Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In *WMT16*, pages 751–758.

Junczys-Dowmunt, Marcin and Roman Grundkiewicz. 2017. An exploration of neural sequence-to-sequence architectures for automatic post-editing. *arXiv preprint arXiv:1706.04138*.

Knowles, R., J. E Ortega, and P. Koehn. 2018. A comparison of machine translation paradigms for use in black-box fuzzy-match repair. In *AMTA 2018*, volume 1, pages 249–255.

Koehn, P. and J. Senellart. 2010. Convergence of translation memory and statistical machine translation. In *AMTA Workshop on MT Research and the Translation Industry*, pages 21–31.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180.

Koehn, Philipp. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

Kranias, L. and A. Samiotou. 2004. Automatic translation memory fuzzy match post-editing: a step beyond traditional TM/MT integration. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 331–334, Lisbon, Portugal.

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

Luong, M. and C. D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*.

Negri, Matteo, Marco Turchi, Rajen Chatterjee, and Nicola Bertoldi. 2018. eSCAPE: a Large-scale Synthetic Corpus for Automatic Post-Editing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, may.

Ortega, J. E., F. Sánchez-Martínez, and M. L. Forcada. 2016. Fuzzy-match repair using black-box machine translation systems: what can be expected? In *AMTA 2016, vol. 1*, pages 27–39.

Ortega, John E, Weiyi Lu, Adam Meyers, and Kyunghyun Cho. 2018. Letting a neural network decide which machine translation system to use for black-box fuzzy-match repair.

Pal, Santanu, 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)*, volume 2, pages 281–286.

Papineni, K., S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL 2002*, pages 311–318.

Parton, Kristen, Nizar Habash, Kathleen McKeown, Gonzalo Iglesias, and Adrià de Gispert. 2012. Can Automatic Post-Editing Make MT More Meaningful? In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 111–118.

Pilevar, Abdol Hamid. 2011. Using Statistical Post-editing to Improve the Output of Rule-based Machine Translation System. *International Journal of Computer Science and Communication*.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Simard, M. and P. Isabelle. 2009. Phrase-based machine translation in a computer-assisted translation environment. *Proceeding of the Twelfth Machine Translation Summit (MT Summit XII)*, pages 120–127.

Simard, Michel, Cyril Goutte, and Pierre Isabelle. 2007. Statistical Phrase-Based Post-Editing. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 508–515.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *AMTA*, volume 200.

Tan, Yiming, Zhiming Chen, Liu Huang, Lilin Zhang, Maoxi Li, and Mingwen Wang. 2017. Neural post-editing based on quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 655–660.

Tebbifakhr, Amirhossein, Ruchit Agrawal, Rajen Chatterjee, Matteo Negri, and Marco Turchi. 2018. Multi-source transformer with combined losses for automatic post editing. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 859–865, Belgium, Brussels, October. Association for Computational Linguistics.

Terumasa, Ehara. 2007. Rule Based Machine Translation Combined with Statistical Post Editor for Japanese to English Patent Translation. In *Proceedings of the XI Machine Translation Summit*, pages 13–18.