The FBK Participation in the WMT 2016 Automatic Post-editing Shared Task

Rajen Chatterjee[1,2], **José G. C. de Souza**[2], **Matteo Negri**[2], **Marco Turchi**[2]
[1]University of Trento, Italy
[2]Fondazione Bruno Kessler, Italy
{chatterjee, desouza, negri, turchi}@fbk.eu

Abstract

In this paper, we present a novel approach to combine the two variants of phrase-based APE (*monolingual* and *context-aware*) by a factored machine translation model that is able to leverage benefits from both. Our factored APE models include part-of-speech-tag and class-based neural language models (LM) along with statistical word-based LM to improve the fluency of the post-edits. These models are built upon a data augmentation technique which helps to mitigate the problem of over-correction in phrase-based APE systems. Our primary APE system further incorporates a quality estimation (QE) model, which aims to select the best translation between the MT output and the automatic post-edit. According to the shared task results, our primary and contrastive (which does not include the QE module) submissions have similar performance and achieved significant improvement of 3.31% TER and 4.25% BLEU (relative) over the baseline MT system on the English-German evaluation set.

1 Introduction

Translation from and to multiple languages is a growing need of this era. Especially in a multilingual continent like Europe this poses a challenge to the language service providers (LSPs) that need to quickly deliver high quality translations. To cope with the increasing demand, the LSPs have shifted human translation from a completely manual process to a semi-automated one, with the help of computer-assisted translation (CAT) tools. CAT tools are indeed becoming a standard and ubiquitous tool for LSPs, which have to daily face the trade-off between quality and productivity, under the pressure of a growing demand. Machines, however, are not yet perfect: machine translation (MT), in particular, is often prone to systematic errors that human post-editing (PE) has to fix before publication. This process of translation results in the generation of parallel data consisting of MT output on one side and its corrected version on the other side. This data can be leveraged to develop Automatic Post-Editing (APE) systems capable not only to spot recurring MT errors, but also to correct them (in a broad sense, ranging from fixing typos to adapting terminology to a specific domain or even modeling the personal style of an individual translator). These capabilities become crucial especially when the MT system used to produce the translation suggestions is a "black-box" whose inner workings are not accessible and can not be tuned or re-trained (a frequent condition for small LSPs).

A recent study on APE by Chatterjee et al. (2015b) over six language pairs have reported consistent improvement (7.3% to 14.7% TER reduction) in the quality of machine translated text across all language pairs. They performed the experiments using the state-of-the-art statistical phrase-based machine translation technique with two variants, which are discussed briefly in Section 2. Based on the observed complementarity between the two variants and the room for mutual improvement, in Section 3 we present a factored APE model capable to leverage the two methods. In Section 4 we describe how to create different representations of the data in order to train each of the variants (monolingual, context-aware) and the factored models. Different configurations of our experiments and their corresponding results are discussed in Section 5. The results of our submissions in the shared task are reported in Section 6, followed by concluding remarks in Section 7.

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 745–750,
Berlin, Germany, August 11-12, 2016.

2 Statistical APE Approaches

Most of the current statistical APE systems follow the phrase-based machine translation approach. They mainly differ in the way the data is represented in the parallel corpus. Unlike MT systems where the parallel corpus is made up of source and target language texts, APE systems use either i) MT text or ii) MT text with source annotations on the source side, and post-edits on the target side of the parallel corpus. The former variant (to use only MT text on the source side) was proposed by Simard et al. (2007), also known as *monolingual* translation, and the latter variant was proposed by Béchara et al. (2011), which is known as *context-aware* translation. The *monolingual* translation approach is more robust, it better generalizes the post-editing rules, and is less prone to word alignment errors which eventually impact on the quality of the post-editing rules. However, since the post-editing rules are learned from (*mt, pe*) (*mt*: machine translated; *pe*: post-edited) pairs, it loses connection with the source sentence, which implies that information lost or distorted in the machine translation process are impossible to recover by the APE system. This issue was addressed by the *context-aware* variant that annotates each word in the machine translated text by the corresponding source word (obtained from word alignment information between the *mt* and *source* text) to form a joint representation (*mt#source*) that represents the new source side of the parallel corpus (as shown in Table 1).

Source	`See Paint on 3D models .`
MT output	`Siehe Bemalen von 3D-Modellen .`
Joint Representation	`Siehe#See Bemalen#Paint_on von#on 3D-Modellen#3D_models .`

Table 1: An example of joint representation used in *context-aware* translation.

APE systems trained with the *context-aware* variant are more precise because they have the power to disambiguate when a *mt* word is a correct translation and when it should be post-edited, by having knowledge of the source context. However, this variant faces two potential problems. First, preserving the source context results in multiple representations of the same *mt* word (each *mt* word can be aligned to multiple *source* words), causing a high increase of the vocabulary size, and, consequently, higher data sparseness that will eventually reduce the reliability of the word alignments and, consequently, of the post-editing rules. Second, the joint representation (*mt#source*) may be affected by the word alignment errors which may mislead the learning of translation options. Moreover, a technical problem with this representation occurs during tuning of the system. Since the input is a joint representation, the OOVs (*mt#source*) penalize the tuning metric even if the *mt* in *mt#source* is a correct translation thereby affecting the tuning process. To address these issues and to leverage the complementarity of the two alternative APE approaches, we propose a more elegant approach that combines them into a factored model as described in the following section.

3 Factored APE model

The factored machine translation model was proposed by Koehn and Hoang (2007). It enables a straightforward integration of additional annotation (called factors) at the word-level. These factors can be linguistic markup or automatically generated word classes. To build our factored APE systems, we pre-process the training data to obtain the factored representation. A fragment of our parallel corpus with factored representation is shown in Table 2. The source side of the parallel corpus has 2 factors (*mt_word* and *source_word*, similar to the joint representation), and the target side contains 3 factors (*pe_word*, *pos-tag*, and *class-id*). In this representation we can define:

- A word alignment mapping between *mt_word*<->*pe_word*. This helps to mitigate the problem of word alignment of *context-aware* APE approach;
- A translation mapping between *mt_word*<->*pe_word* (*monolingual* translation), and *mt_word*|*source_word*<->*pe_word* (*context-aware* translation). This allows us to leverage both the models during decoding;
- A generation mapping between *pe_word*<->*pos-tag*, and *pe_word*<->*class-id*. This allows us to improve the fluency of the trans-

Parallel Corpus	
Source (mt_word|source_word)	**Target** (pe_word|pos-tag|class-id)
Siehe|See Bemalen|Paint_on von|on 3D-Modellen|3D_models .|.	Siehe|ADV|104 "|$(\|373 Bemalen\|NN\|40 von\|APPR\|382 3D-Modellen\|NN\|137 .\|$.|451 "|$(|373
Bildrate|Framerate des|of_the Videos|video MP4|MP4 .|.	Bildrate|NN|339 des|ART|407 MP4-Videos|NN|41 .|$.|451

Table 2: Example of parallel corpus with factored representation.

lations by scoring them with both part-of-speech tag and class-based language models.

Source factors: The factor on the source side of the parallel corpus is obtained following the approach to obtain the joint representation (as described in Section 2) for *context-aware* APE, the only difference is that instead of joint representation (*mt#source*) we now have factored representation (*mt*|*source*) suitable to train factored models.

Target factors: We introduce two target factors to measure fluency of the translations at syntactic and semantic levels, i) POS-tag (∼50 tags) obtained using the TreeTagger (Schmid, 1995), and ii) word-class id (∼500 classes) obtained using *mkcls*[1] tool, which clusters words based on bi-gram contextual similarity. These factors are used to learn generation models ($P(pos\text{-}tag|pe)$ and $P(class\text{-}id|pe)$) to generate corresponding target factors for the test sentence, which are scored by their respective LMs during decoding.

4 Data set and Experimental setup

As defined in the shared task, the training data (English-German) consist of 12K triplets of source (*src*), MT output (*mt*), and human post-edits (*pe*). We split the development data (consisting of 1K triplets) released in this shared task into 400 and 600 triplets (selected randomly) to tune and evaluate our APE systems. We use the *pe* from the training data to build a 5-gram word-based statistical language model using the KENLM toolkit (Heafield, 2011), and 8-gram POS-tag and class-based language model using both KENLM (statistical) and the NPLM (neural) (Vaswani et al., 2013) toolkit. To build the joint representation (*mt#src*) and to obtain source factors (*mt*|*src*), we use the word alignment model trained on *src* and *mt* pairs of the training data by using MGIZA++ (Gao and Vogel, 2008).

To develop the APE systems we use the phrase-based statistical machine translation toolkit MOSES (Koehn et al., 2007) with alignment heuristic set to *"grow-diag-final-and"*, and reordering heuristic to *"msd-bidirectional-fe"*. For building the word alignment models we use MGIZA++ (Gao and Vogel, 2008). For tuning the feature weights we use MERT (Och, 2003) optimizing TER (Snover et al., 2006).

We run case-sensitive evaluation with TER, which is based on edit distance, and BLEU (Papineni et al., 2002), which is based on modified n-gram precision. In addition to the standard evaluation metrics, we also measure precision of our APE system using sentence level TER score as defined in Chatterjee et al. (2015a)

$$\text{Precision} = \frac{\text{Number of Improved Sentences}}{\text{Number of Modified Sentences}}$$

where the "Number of Improved Sentences" consists in all the APE outputs that have lower TER than the corresponding MT output and the "Number of Modified Sentences" consists in all the APE outputs that have TER scores different from the TER of the corresponding MT output.

5 Experiments and Results

Baseline: For internal evaluation we consider the MT system as one of the baselines (an APE system outputting the input sentence), and the two variants of phrase-based APE as described in Section 2. The *monolingual* variant is labeled as APE-1 and the *context-aware* as APE-2. The baseline results reported in Table 4 show that the naive *monolingual* APE system already outperforms the MT baseline by 1.5 BLEU score. However, the low precision of the APE systems indicate that they are prone to over-correction and modifies word-

[1] https://github.com/clab/mkcls

	POS-tag LM			Class-based LM			POS-tag & Class-based LM		
Approach	TER	BLEU	Precision	TER	BLEU	Precision	TER	BLEU	Precision
Statistical	24.20	64.29	63.88	24.28	65.08	67.27	24.22	65.12	70.25
Neural	24.06	65.27	71.85	24.07	65.04	68.92	**24.07**	**65.31**	**72.72**

Table 3: Performance of the Factored APE-2 for various LMs (statistical word-based LM is present in all the experiments by default).

s/phrases which are already correct in the MT output.

Baselines	TER	BLEU	Precision
MT system	24.80	63.07	-
APE-1	**24.73**	**64.55**	**55.55**
APE-2	24.68	64.01	54.01

Table 4: Performance of the APE baselines.

Addressing over-correction: In order to avoid the problem of over-correction (making unnecessary corrections), the APE system should learn to preserve the chunks of the input which are already correct. To this aim, we augmented the parallel corpus with the post-editions (12K) available in the training data. So now our training corpus consist of 12K *mt-pe* or *mt#src-pe* pairs (to learn post-editing rules) and an additional 12K *pe-pe* or *pe#src-pe* pairs (to preserve correct input chunks). Replicating the baseline APE systems with the augmented data showed significant improvements with all the evaluation metrics as reported in Table 5. For this reason, we use the augmented parallel data in all the further experiments. Among the two variants we noticed that the APE-2 gets maximum benefit with an absolute precision improvement of 16.40% (from 54.01% to 70.41%).

	TER	BLEU	Precision
APE-1	24.46	64.74	63.27
APE-2	**24.08**	**64.88**	**70.41**

Table 5: Performance of the APE system with data augmentation technique.

Factored APE models: Both the APE variants have their own strengths and weaknesses as discussed in Section 2. To leverage their complementarity, we use factored translation approach as described in Section 3. Before combining the two variants, we decided to replicate the *context-aware* variant in the factored architecture (since it achieved the best performance as reported in Table 5) with the integration of different target LMs. Along with the 5-gram statistical word-based LM, we study the effect on the performance of the APE system of using an additional 8-gram statistical as well as a neural POS-tag and a class-based LMs. The results are reported in Table 3. It is evident that the neural LM performs better than the statistical ones, and the combination of both POS-tag and class-based neural LM has slightly better precision than the individual neural LMs.

We hence decided to use the neural POS-tag and the class-based LMs along with statistical word-based LM for both the variants (*monolingual* and *context-aware*) in the factored architecture. The translation models of both the variants are used together during decoding with the help of the multiple decoding feature available in the MOSES toolkit (Koehn et al., 2007). The results of this combined factored APE system for various tuning strategies (i) MERT to optimize TER, ii) MERT to optimize BLEU, and iii) MIRA to optimize BLEU are shown in Table 6. Although the TER is almost the same for different tuning strategies, but slight improvement is observed with MIRA in terms of BLEU score.

Optimization	TER	BLEU	Precision
MERT-TER	24.03	65.03	69.71
MERT-BLEU	24.07	65.47	65.67
MIRA-BLEU (Contrastive)	**24.04**	**65.56**	**67.47**

Table 6: Performance of the combined factored model for various tuning configurations.

Factored APE model with quality estimation: To improve the performance of our APE system, we build a sentence-level quality estimation model (Mehdad et al., 2012; Turchi et al., 2014; C. de Souza et al., 2015) to decide whether to select the MT output or our factored APE output (MIRA-BLEU configuration from Table 6). To train the QE model we first extract 79 system-independent features that comprise three different aspects of the QE problem, namely: fluency (e.g. language model perplexity of the whole translation sentence), complexity (e.g. average token length of

the source sentence) and adequacy (e.g. ratio between the number of nouns in the source and translation sentences). These features, obtained with the QuEst feature extractor implementation (Specia et al., 2013) are used to train a regression model that predicts the actual post-editing effort as measured by the TER between the MT-generated translation or the factored APE output and a human post-edited version. The regression model was trained using the extremly randomized trees (Geurts et al., 2006) implementation of scikit-learn library (Pedregosa et al., 2011). This method reached competitive results in sentence-level QE share-tasks in previous years (C. de Souza et al., 2013; C. de Souza et al., 2014). To select the final translation we check if the predicted score of MT output is lower[2] than the predicted score of the APE output by at least k points (threshold). We performed experiments with different threshold values, as reported in Table 7. Using QE with threshold of 5 performs slightly better than the one without QE, so our primary submission is the factored model with QE, whereas, the contrastive one is without QE.

Threshold	TER	BLEU	Precision
1	24.18	65.09	72.13
2	24.15	65.34	70.88
3	24.09	65.51	68.15
4	24.02	65.59	68.94
5 (Primary)	**23.99**	**65.65**	**67.83**
6	24.01	65.64	67.98
Contrastive (w/o QE)	24.04	65.56	67.47
Baseline (MT)	24.80	63.07	-

Table 7: Performance of the APE system with quality estimation for various thresholds.

6 Results of our submissions

The shared task evaluation was on 2,000 unseen samples consisting of *source* and *mt* pairs from the same domain of the training data. Our primary submission is a factored APE system which i) is trained with data augmentation technique, ii) leverages the two statistical phrase-based variants (*monolingual*, and *context-aware*), iii) uses a neural POS-tag and class-based LMs along with the statistical word-based LM, and iv) uses a quality estimation model. Our contrastive submission is similar to primary but without quality estimation.

[2]Lower is better since we are predicting TER scores

According to the shared task results (reported in Table 8) both of our submissions achieves similar performance (with minimal difference in TER) with significant improvement of 3.31% TER and 4.25% BLEU (relative) over the baseline MT system. We also observe that the use of quality estimation in our primary submission did not yield the expected improvements.

	TER	BLEU
Baseline (MT)	24.76	62.11
Baseline (APE)	24.64	63.47
Primary	**23.94**	**64.75**
Contrastive	**23.92**	**64.75**

Table 8: Results of the shared task for our submissions

7 Conclusion

In this system description paper, we discussed the potential strength and weakness of the two phrase-based APE variants (*monolingual* and *context-aware*) and showed that their complementarity can be leveraged by combining them in a factored APE model. Factored models made it possible to integrate several target LMs and study their effect on the performance of the APE systems. From our experiments on LMs, we learn that i) using both the POS-tag, and the class-based LM in the APE system is better than using them in isolation, ii) building these LMs using neural approach is much better than statistical ones, and iii) the best LM combination achieves 0.4 BLEU improvement (from 64.88 to 65.31) over the APE system which do not use these LMs. We also showed that the problem of over-correction in phrase-based APE can be mitigated by our data augmentation technique which showed significant improvement of 0.6 TER, 0.8 BLEU, and 16.40% precision, for *context-aware* variant, over APE system which do not use data augmentation. Performance of our primary and contrastive submissions to the shared task were similar with a significant improvement of 3.31% TER and 4.25% BLEU (relative) over the baseline MT system. However, having a layer of quality estimation in our primary submission did not yield expected improvement.

8 Acknowledgements

This work has been partially supported by the EC-funded H2020 project QT21 (grant agreement no. 645452).

References

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

José G. C. de Souza, Christian Buck, Marco Turchi, and Matteo Negri. 2013. FBK-UEdin participation to the WMT13 Quality Estimation shared-task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 352–358.

José G. C. de Souza, Jesús González-Rubio, Christian Buck, Marco Turchi, and Matteo Negri. 2014. FBK-UPV-UEdin participation in the WMT14 Quality Estimation shared-task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, MD, USA, June.

José G. C. de Souza, Matteo Negri, Elisa Ricci, and Marco Turchi. 2015. Online Multitask Learning for Machine Translation Quality Estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 219–228, Beijing, China, July.

Rajen Chatterjee, Marco Turchi, and Matteo Negri. 2015a. The fbk participation in the wmt15 automatic post-editing shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 210–215.

Rajen Chatterjee, Marion Weller, Matteo Negri, and Marco Turchi. 2015b. 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*, Beijing, China.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning*, 63(1):3–42, April.

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

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of EMNLP-CoNLL*, pages 868–876.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 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.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2012. Match without a Referee: Evaluating MT Adequacy without Reference Translations. In *Proceedings of the Machine Translation Workshop (WMT2012)*, pages 171–180, Montréal, Canada.

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.

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*, pages 311–318.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Mathieu Brucher, Mathieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *Proceedings of the Association for Computational Linguistics SIGDAT-Workshop*.

Michel Simard, 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 (NAACL HLT)*, pages 508–515, Rochester, New York.

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.

Lucia Specia, Kashif Shah, José G. C. de Souza, and Trevor Cohn. 2013. QuEstA translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 79–84.

Marco Turchi, Antonios Anastasopoulos, José G. C. de Souza, and Matteo Negri. 2014. Adaptive Quality Estimation for Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL '14.

Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang. 2013. Decoding with large-scale neural language models improves translation. In *Proceedings of the Empirical Methods in Natural Language Processing*, pages 1387–1392.