USAAR: An Operation Sequential Model for Automatic Statistical Post-Editing

Santanu Pal[1]**, Marcos Zampieri**[1,2]**, Josef van Genabith**[1,2]
[1]Saarland University, Saarbrücken, Germany
[2]German Research Center for Artificial Intelligence (DFKI), Germany
{santanu.pal, marcos.zampieri, josef.vangenabith}@uni-saarland.de

Abstract

This paper presents an automatic post-editing (APE) method to improve the translation quality produced by an English–German (EN–DE) statistical machine translation (SMT) system. Our system is based on Operation Sequential Model (OSM) combined with phrased-based statistical MT (PB-SMT) system. The system is trained on monolingual settings between MT outputs (TL_{MT}) produced by a black-box MT system and their corresponding post-edited version (TL_{PE}). Our system achieves considerable improvement over TL_{MT} on a held-out development set. The reported system achieves 64.10 BLEU (1.99 absolute points and 3.2% relative improvement in BLEU over raw MT output) and 24.14 TER and a TER score of 24.14 (0.66 absolute points and 0.25% relative improvement in TER over raw MT output) in the official test set.

1 Introduction

Translations produced by machine translation (MT) systems have improved substantially over the past few decades. This is particularly noticeable for some language pairs (e.g. English to German and English to French) and for domain specific language (e.g. technical documentation). Texts produced by MT systems are now widely used in the translation and localization industry. MT output is post-edited by professional translators and it has become an important part of the translation workflow. A number of studies confirm that post-editing MT output improves translators' performance in terms of productivity and it may also impact translation quality and consistency (Guerberof, 2009; Plitt and Masselot, 2010; Zampieri and Vela, 2014).

With this respect the ultimate goal of MT systems is to provide output that can be post-edited with the least effort as possible by human translators. One of the strategies to improve MT output is to apply automatic post-editing (APE) methods (Knight and Chander, 1994; Simard et al., 2007a; Simard et al., 2007b). APE methods work under the assumption that some errors in MT systems are recurrent and they can be corrected automatically in a post-processing stage thus providing output that is more adequate to be post-edited. APE methods are applied before human post-editing increasing translators' productivity.

This paper presents a new approach to APE which was submitted by the USAAR team to the Automatic Post-editing (APE) shared task at WMT-2016. Our system combines two models: monolingual phrase-based and operation sequential model with an edit distance based word alignment between an English-German (EN-DE) Machine translation output and the corresponding human post-edited version of German Translation (Turchi et al., 2016).

Usually APE tasks focus on fluency errors produced by the MT system. The most frequent ones are incorrect lexical choices, incorrect word ordering, the insertion of a word, the deletion of a word. For the WMT2016 APE task, in order to automatically post-editing, we adopt operation sequential model (OSM) for SMT to build our Statistical APE (SAPE) System. We inspired from the work of Durrani et al. (2011) and Durrani et al. (2015). Since, in OSM model, the translation and reordering operations are coupled in a single generative story: the reordering decisions may depend on preceding translation decisions and translation decisions may depend on preceding reordering decisions. The model provides a natural re-

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 759–763,
Berlin, Germany, August 11-12, 2016. ©2016 Association for Computational Linguistics

ordering mechanism and deal with both local and long-distance re-orderings consistently.

The remainder of the paper is organized as follows. Section 2 describes our proposed system, in particular PB-SMT coupled OSM model. In Section 3, we outline the data used for experiments and complete experimental setup. Section 4 presents the results of the automatic evaluation, followed by conclusion and future work in Section 5.

2 USAAR APE System

Our APE system is based on operational N-gram sequential model which integrates translation and reordering operations into the phrase-based APE system. Traditional PB-SMT (Koehn et al., 2003) provides a powerful translation mechanism which can directly be modelled to a phrase-based SAPE (PB-SAPE) system (Simard et al., 2007a; Simard et al., 2007b; Pal et al., 2015) using target language MT output (TL_{MT}) and their corresponding post-edited version (TL_{PE}) as a parallel training corpus. Unlike PB-SMT, PB-SAPE also follows similar kind of drawbacks such as dependency across phrases, handling discontinuous phrases etc. Our OSM-APE system is based on phrase based N-gram APE model, however reordering approach is essentially different, it considers all possible orderings of phrases instead of pre-calculated orientations. The model represents the post-edited translation process as a linear sequence of operations such as lexical generation of post-edited translation and their orderings. The translation and reordering decisions are conditioned on n previous translation and reordering decisions. The model also can able to consistently modelled both local and long-range reorderings. Traditional OSM based MT model consists of three sequence of operations:

- Generates a sequence of source and/or target words.
- For reordering operations, inserts gaps as explicit target positions
- Forward and backward jump operations

The sequence operation is based on n-gram model. The probability of a n^{th} operation depends on the $n-1$ preceding operations. The generation of post-edited output (pe) from a given MT sentence (mt), the decoder provides a sequence of hypothesis H: h_1,...,h_n and the APE model estimates the probability $p(mt, pe)$ given in Equation 1, from a sequence of I operations O $(o_1, ...o_I)$ for m amount [1] of context has been used.

$$p(mt, pe) \approx \prod_{i=1}^{I} p(o_i|o_{i-m+1}...o_{i-1}) \qquad (1)$$

The decoder searches best translation in Equation 2 from the model using language model $p_{lm}(pe)$

$$pe^* = argmax_{pe} \frac{p(mt, pe)}{p_{pr}(pe)} \times p_{lm}(pe) \qquad (2)$$

$p_{pr}(pe) \approx \prod_{i=1}^{I} p(w_i|w_{i-m+1}...w_{i-1})$, is the prior probability that marginalize the joint probability $p(mt, pe)$. The model is then represented in a log-linear approach (Och and Ney, 2003) (in Equation 3) that makes it useful to incorporate standard features along with several novel features that improve the accuracy.

$$pe^* = argmax_{pe} \sum_{i=1}^{I} \lambda_i h_i(mt, pe) \qquad (3)$$

where λ_i is the weight associated with the feature $h_i(mt, pe)$: $p(mt, pe)$, $p_{pr}(pe)$ and $p_{lm}(pe)$. Apart from this 8 additional features has been included in the log-linear model:

1. Length penalty: Length of the pe in words
2. Deletion penalty
3. Gap bonus: Total number of gap inserted to produce PE sentence
4. Open gap penalty : Number of open gaps, this penalty controls how quickly gap was closed.
5. Distortion: Distance based reordering which is similar to PB-SMT.
6. Gap distance penalty: The gap between mt and pe sentences generated during the generation process.
7. Lexical features: mt–pe and pe–mt lexical translation probability (Koehn et al., 2003).

[1] We use a 6-gram model trained on SRILM-Toolkit (Stolcke, 2002)

3 Experiment

The effectiveness of the present work is demonstrated by using the standard log-linear PB-SMT model for our phrase based SAPE (PB-SAPE) model. The MT outputs are provided by WMT-2016 APE task (c.f Table 1) are considered as baseline system translation. For building our SAPE system, we experimented with various maximum phrase lengths for the translation model and n–gram settings for the language model. We found that using a maximum phrase length of 10 for the translation model and a 6-gram language model produces the best results in terms of BLEU (Papineni et al., 2002) scores for our SAPE model.

The other experimental settings were concerned with word alignment model between TL_{MT} and TL_{PE} are trained on three different aligners: Berkeley Aligner (Liang et al., 2006), METEOR aligner (Lavie and Agarwal, 2007) and TER (Snover et al., 2006). The phrase-extraction (Koehn et al., 2003) and hierarchical phrase-extraction (Chiang, 2005) are used to build our PB-SAPE and hierarchical phrase-based statistical (HPB-SAPE) system respectively. The reordering model was trained with the hierarchical, monotone, swap, left to right bidirectional (hier-mslr-bidirectional) method (Galley and Manning, 2008) and conditioned on both source and target language. The 5-gram target language model was trained using KenLM (Heafield, 2011). Phrase pairs that occur only once in the training data are assigned an unduly high probability mass (i.e. 1). To compensate this shortcoming, we performed smoothing of the phrase table using the Good-Turing smoothing technique (Foster et al., 2006). System tuning was carried out using Minimum Error Rate Training (MERT) (Och, 2003) optimized with k-best MIRA (Cherry and Foster, 2012) on a held out development set of size 500 sentences randomly extracted from training data. Therefore, all model has been build on 11,500 parallel TL_{MT}–TL_{PE} sentences. After the parameters were tuned, decoding was carried out on the held out development test set ('Dev' in Table 1) as well as test set.

Table 1 presents the statistics of the training, development and test sets released for the English–German APE Task organized in WMT-2016. These data sets did not require any preprocessing in terms of encoding or alignment.

	SEN	Tokens		
		EN	DE-MT	DE-PE
Train	12,000	201,505	210,573	214,720
Dev	1,000	17,827	19,355	19,763
Test	2,000	31,477	34,332	–

Table 1: Statistics of the the WMT-2016 APE Shared Task Data Set. SEN: Sentences, EN: English and DE: German

4 Results

We set various APE system settings for our experiments. We start our experiment with the provide TL_{MT} output, considering as baseline.

In the set of experiments are reported in Table 2, first, three word alignment (one statistical based aligner i.e., Berkeley aligner (Liang et al., 2006) and two edit distance based aligners i.e., METEOR aligner (Lavie and Agarwal, 2007) and TER aligner (Snover et al., 2006)) models are integrated separately within both the PB-SAPE as well as the HPB-SAPE systems. As a result, there are three different PB-SAPE (Experiment 2, 3 and 4 in Table 2) and HPB-SAPE (Experiment 5, 6 and 7 in Table 2) systems.

It is evident from Table 2 that the METEOR aligner is performed better than other two aligners. Therefore, our OSM coupled PB-SAPE model ('OSM' in Table 2) used METEOR based alignment. The experiment result shows that compare to other systems in Table 2, our OSM based model performed better in terms of two evaluation metric BLEU (Papineni et al., 2002) and TER. Evaluation result also shows that both PB-SAPE and HPB-SAPE system performed better over baseline system on development set data. The submitted primary system (OSM in Table 2) achieves 3.06% relative (1.99 absolute BLEU points) improvement over baseline[2] . The system also shows similar improvements is terms of TER evaluation measure.

According to the test set evaluation, our system achieves similar improvements as appeared in development set data. Table 3 shows that, there are two types of baseline systems: (i) *Baseline1* – based on raw MT output and (ii) *Baseline2* – based on Statistical APE (Simard et al., 2007b) (a phrase-based system (Koehn et al., 2007) build

[2] In Table 2, the raw MT output of development set data is considered as MT output of the baseline system.

System		Exp.	BLEU	MET	TER
Baseline	WMT MT-PE	1	65.02	47.79	24.42
PB-SAPE	Berkeley Aligner	2	65.89	48.23	24.51
	METEOR Aligner	3	65.97	48.34	24.36
	TER Aligner	4	65.14	47.85	24.96
HPB-SAPE	Berkeley Aligner	5	66.09	48.31	24.56
	METEOR Aligner	6	66.55	48.58	24.51
	TER Aligner	7	65.19	47.91	24.97
OSM	METEOR Aligner	8	67.01	48.80	24.04

Table 2: Systematic Evaluation on the WMT-2016 APE Shared Task Development Set

using MOSES[3] with default settings). There are two different systems called *OSM_Primary* and *OSM_Constrastive* have been submitted to the WMT-2016 APE shared task. The difference between the two submissions is that the *OSM_Primary* system is tuned with all phrase-based setting parameters including OSM parameters while *OSM_Constrastive* is also tuned with similar parameters but excluding OSM parameters. The tuning process of the OSM parameters is conducted with MERT and optimized with MIRA. Our primary submission obtained a BLEU score of 64.10 (1.99 absolute points and 3.2% relative improvement in BLEU) and a TER score of 24.14 (0.66 absolute points and 0.25% relative improvement in TER) over *Baseline1* system. If we consider *Baseline2* system, our primary submission achieved 0.63 absolute points and 0.99% relative improvement in BLEU and 0.50 absolute points and 0.20% relative improvement in TER.

System	BLEU	TER
Baseline1	62.11	24.76
Baseline2	63.47	24.64
OSM_Primary	**64.10**	**24.14**
OSM_Constrastive	64.00	**24.14**

Table 3: Evaluation on the WMT-2016 APE Shared Task Test Set

5 Conclusion and Future Work

This paper presents the USAAR system submitted in the English–German APE task at WMT-2016. The system demonstrates the crucial role METEOR-based alignment and OSM based SAPE can play in SAPE tasks. The use of statistical aligners in PB-SAPE/HPB-SAPE pipeline successfully improve the APE system, however performances with respect to the translations provided by the baseline are not promising. This is the reason behind use of edit distance-based word alignment into the pipeline. The reason for using OSM model is that, the model tightly couples translation and reordering. Apart from that, the OSM model also considers all possible reorderings instead perform search only on a limited number of pre-calculated orderings. The proposed system, OSM-based SAPE approach, was successful in improving over the PB-SAPE as well as HPB-SAPE performance.

The WMT-2016 APE shared task was a great opportunity to test APE methods that can be later applied in real-word post-editing and computer-aided translation (CAT) tools. We are currently working on implementing the APE methods described in this paper to CATaLog, a recently-developed CAT tool that provides translators with suggestions originated from MT and from translation memories (TM) (Nayek et al., 2015; Pal et al., 2016). In so doing, we aim to provide better suggestions for post-editing and we would like to investigate how this impacts human post-editing performance by carrying out user studies.

Acknowledgments

We would like to thank the WMT-2016 APE shared task organizers for this interesting shared task and for prompt replies to all our inquiries.

The research leading to these results has received funding from the EU FP7 Project EXPERT - the People Programme (Marie Curie Actions), under REA grant agreement no. 317471.

References

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *In Proceedings NAACL-HLT*.

[3] http://www.statmt.org/moses/

David Chiang. 2005. A Hierarchical Phrase-based Model for Statistical Machine Translation. In *Proceedings of ACL.*

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A joint sequence translation model with integrated reordering. In *Proceedings of ACL.*

Nadir Durrani, Helmut Schmid, Alexander M. Fraser, Philipp Koehn, and Hinrich Schtze. 2015. The operation sequence model - combining n-gram-based and phrase-based statistical machine translation. *Computational Linguistics*, 41:185–214.

George Foster, Roland Kuhn, and Howard Johnson. 2006. Phrasetable Smoothing for Statistical Machine Translation. In *Proceedings of EMNLP.*

Michel Galley and Christopher D. Manning. 2008. A Simple and Effective Hierarchical Phrase Reordering Model. In *Proceedings of EMNLP.*

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 WMT.*

Kevin Knight and Ishwar Chander. 1994. Automated Post-Editing of Documents. In *Proceedings of AAAI.*

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-based Translation. In *Proceedings of NAACL-HLT.*

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL.*

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of WMT.*

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by Agreement. In *Proceedings of NAACL-HLT.*

Tapas Nayek, Sudip Kumar Naskar, Santanu Pal, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2015. CATaLog: New approaches to tm and post editing interfaces. In *Proceedings of the NLP4TM Workshop*.

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 ACL.*

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*, September.

Santanu Pal, Marcos Zampieri, Sudip Kumar Naskar, Mihaela Vela, and Josef van Genabith. 2016. CATaLog Online: Porting a Post-editing Tool to the Web. In *Proceedings of LREC.*

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL.*

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.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007a. Statistical Phrase-based Post-Editing. In *In Proceedings of NAACL-HLT.*

Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. 2007b. Rule-Based Translation with Statistical Phrase-Based Post-Editing. In *Proceedings of WMT.*

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 AMTA.*

Andreas Stolcke. 2002. Srilm – an extensible language modeling toolkit. In *IN Proceedings of ICSLP.*

Marco Turchi, Rajen Chatterjee, and Matteo Negri. 2016. WMT16 APE shared task data. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

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 HaCaT Workshop.*