Neural Automatic Post-Editing Using Prior Alignment and Reranking

Santanu Pal[1], **Sudip Kumar Naskar**[2], **Mihaela Vela**[1], **Qun Liu**[3] **and**
Josef van Genabith[1,4]

[1]Saarland University, Saarbrücken, Germany
[2]Jadavpur University, Kolkata, India
[3]ADAPT Centre, School of Computing, Dublin City University, Ireland
[4]German Research Center for Artificial Intelligence (DFKI), Germany
[1]{santanu.pal, josef.vangenabith}@uni-saarland.de
[1]m.vela@mx.uni-saarland.de
[2]sudip.naskar@cse.jdvu.ac.in
[3]qun.liu@dcu.ie

Abstract

We present a second-stage machine translation (MT) system based on a neural machine translation (NMT) approach to automatic post-editing (APE) that improves the translation quality provided by a first-stage MT system. Our APE system (APE_{Sym}) is an extended version of an attention based NMT model with bilingual symmetry employing bidirectional models, $mt \rightarrow pe$ and $pe \rightarrow mt$. APE translations produced by our system show statistically significant improvements over the first-stage MT, phrase-based APE and the best reported score on the WMT 2016 APE dataset by a previous neural APE system. Re-ranking (APE_{Rerank}) of the n-best translations from the phrase-based APE and APE_{Sym} systems provides further substantial improvements over the symmetric neural APE model. Human evaluation confirms that the APE_{Rerank} generated PE translations improve on the previous best neural APE system at WMT 2016.

1 Introduction

The ultimate goal of MT systems is to provide fully automatic publishable quality translations. However, existing MT systems often fail to deliver this. To achieve sufficient quality, translations produced by MT systems often need to be corrected by human translators. This task is referred to as post-editing (PE). PE 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). Nonetheless, translations produced by MT systems have improved substantially and consistently over the last two decades and are now widely used in the translation and localization industry. To enhance the quality of automatic translation without changing the original MT system itself, an additional plug-in post-processing module, e.g. a second stage monolingual MT system (an APE system), can be introduced. This may lead to a more reasonable and feasible solution compared to rebuilding the first-stage MT system. APE can be defined as as an automatic method for improving raw MT output, before performing actual human post-editing (Knight and Chander, 1994). APE assumes the availability of source texts (src), corresponding MT output (mt) and the human post-edited (pe) version of mt. However, APE systems can also be built without the availability of src, by using only sufficient amounts of target side "mono-lingual" parallel mt–pe data. Usually APE tasks focus on systematic errors made by first stage MT systems, acting as an effective remedy to some of the inaccuracies in raw MT output. APE approaches cover a wide methodological range such as SMT techniques (Simard et al., 2007a; Simard et al., 2007b; Chatterjee et al., 2015; Pal et al., 2015; Pal et al., 2016d) real time integration of post-editing in MT (Denkowski, 2015), rule-based approaches to APE (Mareček et al., 2011; Rosa et al., 2012), neural APE (Junczys-Dowmunt and Grundkiewicz, 2016; Pal et al., 2016b), multi-engine and multi-alignment APE (Pal et al., 2016a), etc.

In this paper we present a neural network based APE system to improve raw first-stage MT output

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 349–355,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

quality. Our neural model of APE is based on the work described in Cohn et al. (2016) which implements structural alignment biases into an attention based bidirectional recurrent neural network (RNN) MT model (Bahdanau et al., 2015). Cohn et al. (2016) extends the attentional soft alignment model to traditional word alignment models (IBM models) and agreement over both translation directions (in our case $mt \rightarrow pe$ and $pe \rightarrow mt$) to ensure better alignment consistency. We follow Cohn et al. (2016) in encouraging our alignment models to be symmetric (Och and Ney, 2003) in both translation directions with embedded prior alignments. Different from Cohn et al. (2016), we employed prior alignment computed by a hybrid multi-alignment approach. Evaluation results show consistent improvements over the raw first-stage MT system output and over the previous best performing neural APE (Junczys-Dowmunt and Grundkiewicz, 2016) on the WMT 2016 APE test set. In addition we show that re-ranking n-best output from baseline and enhanced PB-SMT APE systems (Section 3) together with our neural APE output provides further statistically significant improvements over all the other systems.

The main contributions of our research are (i) an application of bilingual symmetry of the bidirectional RNN for APE, (ii) using a hybrid multi-alignment based approach for the prior alignments, (iii) a smart way of embedding word alignment information in neural APE, and (iv) applying reranking for the APE task.

The remainder of the paper is structured as follows: Section 2 describes the our symmetric neural APE model. Section 3 describes the experimental setup and presents the evaluation results. Section 4 summarizes our work, draws conclusions and presents avenues for for future work.

2 Symmetric Neural Automatic Post Editing Using Prior Alignment

Below we describe bilingual symmetry of bidirectional RNN with embedded prior word alignment for APE.

2.1 Hybrid Prior Alignment

The monolingual *mt–pe* parallel corpus is first word aligned using a hybrid word alignment method based on the alignment combination of three different statistical word alignment methods: (i) GIZA++ (Och, 2003) word alignment with grow-diag-final-and (GDFA) heuristic (Koehn, 2010), (ii) Berkeley word alignment (Liang et al., 2006), and (iii) SymGiza++ (Junczys-Dowmunt and Szał, 2012) word alignment, as well as two different edit distance based word aligners based on Translation Edit Rate (TER) (Snover et al., 2006) and METEOR (Lavie and Agarwal, 2007). We follow the alignment strategy described in (Pal et al., 2013; Pal et al., 2016a). The aligned word pairs are added as additional training examples to train our symmetric neural APE model. Each word in the first stage MT output is assigned a unique id (sw_{id}). Each *mt–pe* word alignment also gets a unique identification number (a_{id}) and a vector representation is generated for each such a_{id}. Given a sw_{id}, the neural APE model is trained to generate a corresponding a_{id} based on the context sw_{id} appears in. The APE words are generated from a_{id} by looking up the hybrid prior alignment look-up table (LUT). Neural MT jointly learns alignment and translation. Replacing the source and target words by sw_{id} and a_{id}, respectively, implicitly integrates the prior alignment and lessens the burden of the attention model. Secondly, our approach bears a resemblance to the sense embedding approach (Li and Jurafsky, 2015) since an embedding is generated for each (sw_{id}, a_{id}) pair.

2.2 Symmetric Neural APE

Our symmetric neural APE model (APE_{Sym}) is inspired by the bilingual symmetry (Cohn et al., 2016) of the bidirectional RNN based MT (Bahdanau et al., 2015). Bilingual symmetry inferences of both directional attention models are combined. The bidirectional RNN is based on an *encoder-decoder* architecture, where the first-stage MT output is encoded into a distributed representation, followed by a decoding step which generates the APE translation. The *encoder* consists of a forward RNN ($h_i^{\rightarrow} = f(h_{i-1}^{\rightarrow}, r_i)$), which reads in each input string **x** sequentially from x_1 to x_m at each time step i, and a backward RNN ($h_t^{\leftarrow} = f(h_{i+1}^{\leftarrow}, r_i)$), which reads in the opposite direction, i.e., sequentially from x_m to x_1, f being an activation function, defined as an elementwise logistic sigmoid with an LSTM unit. Here, $r_i = \sigma(W^r \bar{E} x_i + U^r h_{i-1})$, where $\bar{E} \in R^{m \times k_x}$ is the word embedding matrix of the MT output, $W^r \in R^{m \times n}$ and $U^r \in R^{n \times n}$ are weight matrices, m is the word embedding dimensionality and n represents the number of hidden units.

k_x and k_y correspond to the vocabulary sizes of source and target languages, respectively. The hidden state of the *decoder* at time t is computed as $\eta_t = f(\eta_{t-1}, y_{t-1}, c_t)$, where c_t is the context vector computed as $c_t = \sum_{i=1}^{T_x} \alpha_{ti} h_i$. Here, α_{ti} is the weight of each h_i and can be computed as in Equation 1

$$\alpha_{ti} = \frac{exp(e_{ti})}{\sum_{j=1}^{m} exp(e_{tj})} \quad (1)$$

where $e_{ti} = a(\eta_{t-1}, h_i)$ is a word alignment model. Based on the input (mt) and output (pe) sequence lengths, T_x and T_y, the alignment model is computed $T_x \times T_y$ times as in Equation 2

$$a(\eta_{t-1}, h_i) = v^{aT} \tanh(W^a \eta_{t-1} + U^a h_i) \quad (2)$$

where $W^a \in \mathbf{R}^{m \times n}$, $U^a \in \mathbf{R}^{n \times 2n}$ and $v^a \in \mathbf{R}^n$ are the weight matrices of n hidden units. T denotes the transpose of a matrix. Each hidden unit η_t can be defined in Equation 3

$$\eta_t = \tanh(W^d E y_{t-1} + U^d \eta_{t-1} r_t + C c_t) \quad (3)$$

where, $r_t = \sigma(W^r E y_{t-1} + U^r \eta_{t-1} + C^r c_t)$ E is the word embedding matrix for PE. $W^d, W^r \in R^{n \times m}, U^d, U^r \in R^{n \times n}$ and $C, C^r \in R^{n \times 2n}$ are weights. The joint training of the bilingual symmetry models is established using symmetric training with trace bonus, which is computed as $-t_r(\alpha^{mt \to pe} \alpha^{pe \to mt^T})$. This involves optimizing L as in Equation 4.

$$L = -\log p(pe|mt) - \log p(mt|pe) + \gamma B \quad (4)$$

where B links the two models as $B = sum_j \sum_i \alpha_{i,j}^{mt \to pe} \alpha_{j,i}^{pe \to mt}$, where α are alignment (attention) matrices of $T_x \times T_y$ dimensions. The advantage of symmetrical alignment cells is that they are normalized using softmax (values in between 0 and 1), therefore, the trace term is bounded above by $min(T_x, T_y)$, representing perfect one-to-one alignments in both directions.

To train each directional attention model ($mt \to pe$ and $pe \to mt$), we follow the work described in Cohn et al. (2016), where absolute positional bias between the MT and PE translation (as in IBM Model 2), fertility relative position bias (as in IBM Models 3, 4, 5) and HMM-based Alignment (Vogel et al., 1996) are incorporated with an attention based soft alignment model.

3 Experiments and Results

We carried out our experiments on the 12K English–German WMT 2016 APE task training data described in Bojar et al. (2016) and for some experiments we also use the 4.5M artificially developed APE data described in Junczys-Dowmunt and Grundkiewicz (2016). The training data consists of English–German triplets containing source English text (src) from the IT domain, corresponding German translations (mt) from a first-stage MT system and the corresponding human post-edited version (pe). Development and test data contain 1,000 and 2,000 triplets respectively.

We considered two baselines: (i) the raw MT output provided by the first-stage MT system serves as *Baseline1* (WMT_{B_1}) and (ii) *Baseline2* (WMT_{B_2}) is based on Statistical APE, a second-stage phrase-based SMT system (Koehn et al., 2007) built using MOSES[1] with default settings and trained on the 12K mt–pe training data.

In addition to the two baselines, we also compared our attention based neural mt–pe symmetric model (APE_{Sym}) against the best performing system (WMT_{Best}) in the WMT 2016 APE task and the standard log-linear mt–pe PB-SMT model with hybrid prior alignment as described in Section 2.1 (APE_{B2}). APE_{B2} and APE_{Sym} models are trained on 4.55M (4.5M + 12K + pre-aligned word pairs) parallel mt–pe data. The pre-aligned word pairs are obtained from the hybrid prior word alignments (Section 2.1) of the 12K WMT APE training data. For building our APE_{B2} system, we set a maximum phrase length of 7 for the translation model, and a 5-gram language model was trained using KenLM (Heafield, 2011). Word alignments between the mt and pe (4.5M synthetic mt-pe data + 12K WMT APE data) were established using the Berkeley Aligner (Liang et al., 2006), while word pairs from hybrid prior alignment (Section 2.1) between mt–pe (12K data) were used for the additional training data to build APE_{B2}. 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 the source and target language. Phrase pairs that occur only once in the training data are assigned an unduly high probability mass (1) in the PB-SMT framework. 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).

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

For setting up our neural network, previous to training the APE_{Sym} model, we performed a number of preprocessing steps on the mt–pe parallel training data. First, we prepare a LUT containing mt–pe hybrid prior word alignment above (Section 2.1) a certain lexical translation probability threshold (0.3). To ensure efficient use of the hybrid prior alignment we replaced each mt word by a unique identification number (sw_{id}) and each pe word by a unique alignment identification number (a_{id}) (cf. Section 2.1). Afterwards, to effectively reduce the number of unknown words to zero, we follow a preprocessing mechanism similar to Junczys-Dowmunt and Grundkiewicz (2016). We built our APE_{Sym} model with a single-layer LSTM as encoder and two-layer LSTM as decoder, using 1024 embedding, 1024 hidden and 512 alignment dimensions. Our neural APE model is trained end-to-end using stochastic gradient descent (SGD), allowing up to 20 epochs. The development set was used for regularization by early stopping, which terminated training after 10 epochs. The APE_{Sym} model maintains bilingual symmetry, and the inferences of both directional models are combined. In a bid to further improve the translation quality, we also preformed re-ranking (cf. APE_{Rerank} in Table 1). For re-ranking[2], we generated 100-best translations from each participating system (WMT_{B2} and APE_{B2}) along with our APE_{Sym} model. As with the SMT based APE output, we added log probability features from our neural models. Additionally, we used the following features: n-gram ($n = 3...7$) language model probability as well as perplexity normalized by sentence length, minimum Bayes risk scores, and mt–pe length ratio. We trained the re-ranking model on the development set using MERT with 100-distinct best translations of each participating system which are optimized on BLEU.

3.1 Automatic Evaluation

Table 1 provides a comparison of the baseline WMT_{B_1}, WMT_{B_2}, WMT_{Best}, APE_{B2}, APE_{Sym} and the APE_{Rerank} system. Automatic evaluation was carried out in terms of BLEU (Papineni et al., 2002), METEOR and TER. Some general trends can be observed across all metrics. Automatic post-editing, even trained on a small amount of training data (WMT_{B_2}), provides improvements over raw MT output in general. Additional training data, even artificially generated, helps improve system performance (compare APE_{B2} with WMT_{B_2}). Neural MT performs better than PB-SMT based approaches for the post-editing task on large amounts of training data (compare WMT_{Best} and APE_{B2} with WMT_{B_2}). Our APE_{Sym} system based on Cohn et al. (2016) with hybrid embedded prior word alignment provides the best performance among all the individual APE systems and surpasses the WMT_{Best} system. The APE_{Rerank} system performs significantly better than all the individual systems. The scores marked with * in Table 1 indicate statistically significant improvements ($p < 0.01$) as measured by bootstrap resampling (Koehn, 2004) over the corresponding score in the previous row. We observed that APE_{Sym} contributed to the majority (70.65%) of the translations selected by APE_{Rerank}.

3.2 Human Evaluation

In order to assess the performance of the APE system, we conducted experiments with human evaluators comparing our best APE system (APE_{Rerank}) against the WMT 2016 winning APE system (WMT_{Best}). Human evaluation was carried out using *CATaLog Online*[3] – an online CAT tool (Pal et al., 2016c). Our human evaluators were 18 undergraduate students enrolled in a Translation Studies programme, attending a translation technologies class, including sessions on MT and MT evaluation. All students were native speakers of German with at least a B2 level of English. During evaluation students were presented an English source sentence and two German MT outputs (APE_{Rerank} and WMT_{Best}), the ordering of the MT outputs being alternated for each presentation. They had to decide between the two MT outputs by marking the translation they consider of better quality in terms of both adequacy and fluency. Each student received a set of 30 sentences for evaluation, with 20 sentences drawn randomly and 10 sentences being common to all students, allowing us to compare the distribution of decisions across all sentences and the 10 common sentences. The outcome of the evaluation is presented in Table 2. Assessors preferred the MT output produced by APE_{Rerank} in 58.5% cases

[2] Our approach is inspired by Och et al. (2004).

[3] http://santanu.appling.uni-saarland.de/CATaLog/

System	Data	BLEU↑	METEOR↑	TER↓
WMT_{B_1}	-	62.11	72.2	24.76
WMT_{B_2}	12K	63.47*	73.3*	24.64*
APE_{B2}	5.1M	64.40*	73.7*	24.10*
WMT_{Best}	5.1M	67.65*	76.1*	21.52*
APE_{Sym}	5.1M	67.87*	76.3*	21.07*
APE_{Rerank}	5.1M	**69.90***	77.5*	**20.70***

Table 1: Automatic evaluation results

and chose the WMT_{Best} output for rest of the cases (i.e., 41.5%). On the 10 common sentences evaluated by all the evaluators, the results show a similar trend (57.8% in favour of APE_{Rerank}, 42.2% for WMT_{Best}).

	540 sentences	180 sentences
APE_{Rerank}	58.5%	57.8%
WMT_{Best}	41.5%	42.2%

Table 2: Selection of suggestions by assessors for all sentences and for only the common sentences.

4 Conclusions and Future Work

In this paper we presented a neural APE model that extends the attention based NMT model to traditional word alignment models and utilizes agreement of bidirectional models for alignment symmetry. The attentions are encouraged to symmetrization in both translation directions. To the best of our knowledge this is the first work on integrating hybrid prior alignment into NMT. Evaluation results show significant improvements over the first-stage raw MT system. Although the APE_{Sym} system provided only small (but significant) improvements over WMT_{Best} system, re-ranking of the n-best outputs of the multiple APE engines yields large improvements. Human evaluation also revealed the superiority of the APE_{Rerank} system over the WMT_{Best} system.

As future work we plan to integrate source knowledge into the neural APE framework. We will also study further the use of standard word alignment information to influence the attention mechanism in neural APE.

Acknowledgments

We would like to thank all the anonymous reviewers for their feedback. 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, MeitY, Government of India, under the Young Faculty Research Fellowship of the Visvesvaraya PhD Scheme for Electronics & IT. Qun Liu and Josef van Genabith is supported by funding from the European Union Horizon 2020 research and innovation programme under grant agreement no 645452 (QT21).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany.

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, Beijing, China.

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating Structural Alignment Biases into an Attentional Neural Translation Model. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 876–885, San Diego, California.

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

George Foster, Roland Kuhn, and Howard Johnson. 2006. Phrasetable smoothing for statistical machine translation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 53–61, Stroudsburg, PA, USA.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 848–856, Stroudsburg, PA, USA.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 187–197, Stroudsburg, PA, USA.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In *Proceedings of the First Conference on Machine Translation*, pages 751–758, Berlin, Germany.

Marcin Junczys-Dowmunt and Arkadiusz Szał, 2012. *SyMGiza++: Symmetrized Word Alignment Models for Statistical Machine Translation*, pages 379–390. Springer Berlin Heidelberg, Berlin, Heidelberg.

Kevin Knight and Ishwar Chander. 1994. Automated Postediting of Documents. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (Vol. 1)*, AAAI '94, pages 779–784, Seattle, Washington, USA.

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 the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180, Prague, Czech Republic.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain.

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

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, Prague, Czech Republic.

Jiwei Li and Dan Jurafsky. 2015. Do Multi-Sense Embeddings Improve Natural Language Understanding? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1722–1732, Lisbon, Portugal.

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, New York, New York.

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, Edinburgh, Scotland.

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

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 161–168, Boston, Massachusetts, USA.

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, Sapporo, Japan.

Santanu Pal, Sudip Naskar, and Sivaji Bandyopadhyay. 2013. A Hybrid Word Alignment Model for Phrase-Based Statistical Machine Translation. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, pages 94–101, Sofia, Bulgaria.

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, and Josef van Genabith. 2016a. Multi-Engine and Multi-Alignment Based Automatic Post-Editing and its Impact on Translation Productivity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2559–2570, Osaka, Japan.

Santanu Pal, Sudip Kumar Naskar, Mihaela Vela, and Josef van Genabith. 2016b. 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, Berlin, Germany.

Santanu Pal, Sudip Kumar Naskar, Marcos Zampieri, Tapas Nayak, and Josef van Genabith. 2016c. CATaLog Online: A Web-based CAT Tool for Distributed Translation with Data Capture for APE and Translation Process Research. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 98–102, Osaka, Japan.

Santanu Pal, Marcos Zampieri, and Josef van Genabith. 2016d. USAAR: An Operation Sequential Model for Automatic Statistical Post-Editing. In *Proceedings of the First Conference on Machine Translation*, pages 759–763, Berlin, Germany.

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, Philadelphia, Pennsylvania.

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, Stroudsburg, PA, USA.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007a. Statistical Phrase-Based Post-Editing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 508–515, Rochester, New York.

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, Prague, Czech Republic.

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, Cambridge, Massachusetts, USA.

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

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based Word Alignment in Statistical Translation. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pages 836–841, Copenhagen, Denmark.