USAAR-SAPE: An English–Spanish Statistical Automatic Post-Editing System

Santanu Pal[1], Mihaela Vela[1], Sudip Kumar Naskar[2], Josef van Genabith[1]
[1]Saarland University, Saarbrücken, Germany
[2]Jadavpur University, Kolkata, India
{santanu.pal, josef.vangenabith}@uni-saarland.de
m.vela@mx.uni-saarland.de
sudip.naskar@cse.jdvu.ac.in

Abstract

We describe the USAAR-SAPE English–Spanish Automatic Post-Editing (APE) system submitted to the APE Task organized in the Workshop on Statistical Machine Translation (WMT) in 2015. Our system was able to improve upon the baseline MT system output by incorporating Phrase-Based Statistical MT (PBSMT) technique into the monolingual Statistical APE task (SAPE). The reported final submission crucially involves hybrid word alignment. The SAPE system takes raw Spanish Machine Translation (MT) output provided by the shared task organizers and produces post-edited Spanish text. The parallel data consist of English Text, raw machine translated Spanish output, and their corresponding manually post-edited versions. The major goal of the task is to reduce the post-editing effort by improving the quality of the MT output in terms of fluency and adequacy.

1 Introduction

In this paper, we present the submission of Saarland University (USAAR) to the WMT2015 APE task. The system combines a hybrid word alignment system implementation with a monolingual PBSMT for the language pair English-Spanish (EN-ES), translating from English into Spanish.

In order to achieve the desired translation quality, translations provided by MT systems need to be corrected by human translators. Automatic MT post-editing (APE) (Knight and Chander, 1994) is the method of improving raw MT output, before performing human post-editing on it. The objective is to decreases the amount of errors produced by the MT systems, achieving in the end a productivity increase in the translation process.

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 WMT2015 APE task, we adapted our system in order to automatically post-edit lexical choice errors, word insertions and deletions. The method is also able to correct to some extent word ordering.

The remainder of the paper is organized as follows. Section 2 gives an overview of the related work, Section 3 describes the various components of our system, in particular the corpus preprocessing module, the hybrid word alignment module and the PBSMT model. In Section 4, we outline the complete experimental setup. Section 5 presents the results of the automatic and human evaluation, followed by conclusion in Section 6.

2 Related Work

In order to implement the correction of repetitive errors in the MT output, various automatic or semi-automatic post-processing or automatic PE techniques have been developed. Although MT output needs to be post-edited by humans to produce publishable quality translation (Roturier, 2009; TAUS/CNGL Report, 2010), it is faster and cheaper to post-edit MT output than to perform human translation from scratch. In some cases, recent studies have shown that the quality of MT output plus PE can exceed the quality of human translation (Fiederer and O'Brien, 2009; Koehn, 2009; De Palma and Kelly, 2009) as well as the productivity (Zampieri and Vela, 2014). Aimed at cost-effective and timesaving use of MT, the PE process needs to be further optimised (TAUS/CNGL Report, 2010). Post-editing can be also used as a MT evaluation method, implying at least source and target language skills, different from ranking, that does nor require specific skills, a homogeneous group of evaluators be-

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 216–221,
Lisboa, Portugal, 17-18 September 2015.

ing enough to perform the task (Vela and van Genabith, 2015) .

The aim of automatic post-editing (APE) is to improve the output of MT by post-processing it. One of the first approaches was the one introduced by Chen and Chen (1997) who proposed a combination of rule-based MT (RBMT) and statistical MT (SMT) systems aiming at merging the positive properties of each system type for a better machine translation output.

Simard et al. (2007a) and Simard et al. (2007b) have shown how a PBSMT system can be used for automatic post-editing of an RBMT system for translations from English to French and French to English. Because RBMT systems tend to produce repetitive errors, they train a SMT system to correct errors, with the aim of reducing the post-editing effort. The SMT system trains on the output of the RBMT system as the source language and the reference human translations as the target language. The evaluation of their system shows that the post-edited output had a better quality than the output of the RBMT system as well as the output of the same SMT system used in standalone translation mode.

Lagarda et al. (2009) use an approach similar to Simard et al. (2007a) for translations from English to Spanish. The evaluation of the method was performed automatically and manually by comparing the APE output with the output from an RBMT system and a SMT system. The two corpora used in the evaluation were transcriptions of parliamentary speeches and medical protocols. The evaluation results have shown that on transcriptions of parliamentary speeches the method improves the RBMT system.

Rosa et al. (2012) and Mareček et al. (2011) applied APE on English-to-Czech MT outputs on morphological level. Based on word alignment, the method learns during the training phase 20 hand-written rules based on the most frequent errors encountered in translation. The method addresses fluency in translation and corrects morphosyntactic categories of a word such as number, gender, case, person and dependency label.

Parton et al. (2012) present an approach to APE consisting of three stages: detecting errors, suggesting and ranking corrections for the errors, and applying the developed suggestions. For the last stage of their method, applying the corrections, Parton et al. (2012) developed two different methodologies, a rule-abased APE and a feedback APE. The rule-based APE performs either insertions or replacement to address an identified error. The feedback APE, an approach similar to the one proposed by Parton and McKeown (2010), passes the possible correction to the MT system, letting the MT decoder decide whether the errors should be corrected and about the method of correcting it. Parton et al. (2012) evaluated their approach with human evaluators and found that the adequacy of post-edited MT output improved both for rule-based and feedback APE. In terms of fluency the human evaluation has shown that adequacy increase in feedback APE is related to fluency but not for rule-based APE.

Denkowski (2015) has developed a method for integrating in real time post-edited MT output into a translation model, by extracting for each input sentence a grammar. The method, based on Levenberg et al. (2010) and Lopez (2008), allows the indexing of the the source and post-edited MT output, as well as the union of the already existing sentence pairs with the new post-edited data. The system can also remember the rules that are consistent with the post-edited data. This way, rules learned from human corections can be preferred. The experiments Denkowski (2015) ran on from English into and out of Spanish and Arabic data show that the process of translating with an adaptive grammar improves performance on post-editing tasks.

3 System Description

Our system is designed with three basic components: corpus preprocessing, hybrid word alignment and a PBSMT system integrated with the hybrid word alignment. The hybrid word alignment consists of the combination of multiple word alignments into a single word alignment table which is later used in a phrase-based SMT (PBSMT) system. Our SMT based SAPE systems were trained on monolingual Spanish MT output and the manually post-edited output.

3.1 Corpus Preprocessing

For training our system we used the sentence aligned training data provided by the organizers of the WMT2015 APE task. The training data consist of 11,272 parallel segments of English to Spanish MT translations as well as the post-edited translations of the MT output. The English source text,

the machine translated Spanish output and the corresponding post-edited version contain 238,335, 257,644 and 257,881 tokens respectively.

The preprocessing of the training corpus was carried out first by stemming the Spanish MT output and the PE data using Freeling (Padró and Stanilovsky, 2012).

3.2 Hybrid Word Alignment

3.2.1 Statistical Word Alignment

GIZA++ (Och and Ney, 2003) is a statistical word alignment tool which implements maximum likelihood estimators for all the IBM-1 to IBM-5 models, a HMM alignment model as well as the IBM-6 model covering many to many alignments. GIZA++ facilitates fast development of statistical machine translation (SMT) systems. Like GIZA++, the Berkley Aligner (Liang et al., 2006) is also used to align words across sentence pairs. The Berkeley word aligner uses an extension of Cross Expectation Maximization and is jointly trained with HMM models. We use a third statistical word aligner called SymGiza++ (Junczys-Dowmunt and Szał, 2012), which modifies the counting phase of each model of Giza++ allowing for updating the symmetrized models between the chosen iterations of the original training algorithms. It computes symmetric word alignment models with the capability of taking advantage of multi-processor systems.

3.2.2 Edit Distance-Based Word Alignment

We use two different kind of edit distance based word aligners, where alignment is based on TER (Translation Edit Rate) and the METEOR word aligner. TER (Snover et al., 2006) was developed for automatic evaluation of MT outputs. TER can align two strings such as the reference (in this case the PE translation) and the hypothesis (MT output). In the our work, the reference string has been chosen to be the confusion network skeleton, and the hypotheses are aligned independently using the skeleton. These pair-wise alignments may be consolidated to form a confusion network. TER measures the ratio between the number of edit operations that are required to turn a hypothesis *H* into the corresponding reference *R* to the total number of words in the *R*. The allowable edit types include insertion (Ins), substitution (Sub), deletion (Del) and phrase shifts (Shft). TER is computed as

$$TER(H,R) = \frac{(Ins + Del + Sub + Shft) * 100\%}{\textit{total number of words in R}} \quad (1)$$

METEOR Alignment (Lavie and Agarwal, 2007) is also an automatic MT evaluation metric which provides an alignment between hypothesis (here the MT output) and reference (here the PE translation). Given a pair of strings such as *H* and *R* to be compared, METEOR initially establishes a word alignment between them. The alignment is provided by a mapping method between the words in the hypothesis *H* an reference *R* transaltion, which is built incrementally by the following sequence of word-mapping modules:

- **Exact:** maps if they are exactly the same
- **Porter stem:** maps if they are the same after they are stemmed using the Porter stemmer
- **WN synonymy:** maps if they are considered synonyms in WordNet

If multiple alignments exist, METEOR selects the alignment for which the word order in the two strings is most similar (i.e. having fewest crossing alignment links). The final alignment is produced between *H* and *R* as the union of all stage alignments (e.g. exact, Porter stemming and WN synonymy).

3.2.3 Hybridization

The hybrid word alignment method combines two different kinds of word alignment: the statistical alignment tools such as GIZA++ word alignment with grow-diag-final-and (GDFA) heuristic (Koehn, 2010) and SymGiza++ (Junczys-Dowmunt and Szał, 2012) and the Berkeley aligner (Liang et al., 2006), as well as edit distance-based aligners (Snover et al., 2006; Lavie and Agarwal, 2007). In order to combine these different word alignment tables (Pal et al., 2013) we used a mathematical union method. For the union method, we hypothesise that all alignments are correct. Duplicate entries are removed.

3.3 Phrase-Based SMT

Translation is modelled in SMT as a decision process, in which the translation

$$e_1^L = e_1 ... e_i ... e_I \quad (2)$$

of a source sentence

$$f_1^J = f_1 ... f_j ... f_J \quad (3)$$

is chosen to maximize in equation (4):

$$argmax_{I,e_1^L} P(e_1^L|f_1^J) = \quad (4)$$
$$argmax_{I,e_1^L} P(f_1^J|e_1^L) * P(e_1^L)$$

where $P(f_1^J|e_1^L)$ is the translation model and $P(e_1^L)$ the target language model. In log-linear phrase-based SMT, the posterior probability is directly modeled as a log-linear combination of features (Och and Ney, 2003), involving *M* translational features, and the language model, as in equation (5):

$$logP(e_1^L|f_1^J) = \quad (5)$$
$$\sum_{m=0}^{M} \lambda_m h_m(f_1^J, e_1^L, s_1^k) + \lambda_{LM} logP(e_1^L)$$

where $s_1^k = s_1 \ldots s_k$ denotes a segmentation of the source and target sentences respectively into the sequences of phrases ($\hat{e}_1^k = \hat{e}_1 \ldots \hat{e}_k$) and ($\hat{f}_1^k = \hat{f}_1 \ldots \hat{f}_k$) such that (we set $i_0 = 0$) in equation (6):

$$\forall 1 \leq k \leq K, s_k = (i_k, b_k, j_k), \quad (6)$$
$$\hat{e}_k = e_{i_{k-1}+1} ... e_{i_k}, \hat{f}_k = f_{b_k} ... f_{j_k}$$

and each feature $\hat{h}_m$ in (5) can be rewritten as in (7):

$$h_m(f_1^J, e_1^L, s_1^k) = \sum_{k=1}^{K} \hat{h}_m(\hat{f}_k, \hat{e}_k, s_k) \quad (7)$$

where $\hat{h}_m$ is a feature that applies to a single phrase-pair. It thus follows (8):

$$\sum_{m=1}^{M} \lambda_m \sum_{k=1}^{K} \hat{h}_m(\hat{f}_k, \hat{e}_k, s_k) = \sum_{k=1}^{K} \hat{h}(\hat{f}_k, \hat{e}_k, s_k) \quad (8)$$

where $\hat{h} = \sum_{k=1}^{K} \lambda_m \hat{h}_m$.

4 Experiments

We performed experiments on the development set provided by the organizers of the APE task in the WMT2015.

4.1 Data

Table 1 presents the statistics of the training, development and test sets released for the English–Spanish SAPE Task organized in WMT'2015. These data sets did not require any preprocessing in terms of encoding or alignment.

	SEN	Tokens		
		EN	ES-MT	ES-PE
Train	11,272	238,335	257,644	257,881
Dev	1,000	21,617	23,213	23,098
Test	1,817	38,244	40,925	–

Table 1: Statistics. SEN: Sentences, EN: English and ES: Spanish

4.2 Experimental Settings

The effectiveness of the present work is demonstrated by using the standard log-linear PBSMT model. 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 7 for the translation model and a 5-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 hybrid word alignment training algorithms (described in Section 3) and the phrase-extraction (Koehn et al., 2003). The reordering model was trained with the hierarchical, monotone, swap, left to right bidirectional (hier-mslr-bidirectional) (Galley and Manning, 2008) method 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 alleviate 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) optimised with k-best MIRA (Cherry and Foster, 2012) on a held out development set. After the parameters

were tuned, decoding was carried out on the held out test set.

5 Evaluation

The evaluation of our SAPE system was performed on the 1817 Spanish sentences. The baseline consisted of two systems, an MT baseline system and the APE the system of (Simard et al., 2007a). The evaluation was carried out using HTER (TER with human targeted references) score. In this year's WMT seven groups made a submission to the APE task. From the seven systems, our system was ranked on the third place, achieving a HTER score of 23.426 for case sensitive evaluation and 22.710 for the case insensitive evaluation, outperforming the baseline APE system scoring 23.839 for the case sensitive evaluation and 23.130 for the case insensitive evaluation.

6 Conclusion

This paper presents our system submitted in the English–Spanish APE Task for WMT2015. The system demonstrates the crucial role hybrid word alignment can play in SAPE tasks. Edit-distance based monolingual aligner provides alignment for our SAPE system. Incorporating hybrid word alignment into the state-of-the-art PBSMT pipeline provides additional improvements over the baseline APE system.

Acknowledgments

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

Kuang-Hua Chen and Hsin-Hsi Chen. 1997. A Hybrid Approach to Machine Translation System Design. *Computational Linguistics and Language Processing*, 23:241–265.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *In Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies (NAACL-HLT)*, pages 427–436.

Donald De Palma and Nataly Kelly. 2009. Project Management for Crowdsourced Translation: How User-Translated Content Projects Work in Real Life. *Translation and Localization Project Management: The Art of the Possible*, pages 379–408.

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

Rebecca Fiederer and Sharon O'Brien. 2009. Quality and Machine Translation: a Realistic Objective. *Journal of Specialised Translation*, 11:52–74.

George Foster, Roland Kuhn, and Howard Johnson. 2006. Phrasetable Smoothing for Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 53–61.

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)*, pages 848–856.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the 6th Workshop on Statistical Machine Translation (WMT)*, pages 187–197.

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

Kevin Knight and Ishwar Chander. 1994. Automated Post-Editing of Documents. In *Proceedings of the 12th National Conference on Artificial Intelligence*, pages 779–784.

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

Philipp Koehn. 2009. A Process Study of Computer Aided Translation. *Machine Translation*, 23(4):241–263.

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

Antonio Lagarda, Vicent Alabau, Francisco Casacuberta, Roberto Silva, and Enrique Díaz-de Liaño. 2009. Statistical Post-Editing of a Rule-based Machine Translation System. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies (NAACL-HLT)*, pages 217–220, Stroudsburg, PA, USA.

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 2nd Workshop on Statistical Machine Translation (WMT)*, pages 228–231.

Abby Levenberg, Chris Callison-Burch, and Miles Osborne. 2010. Stream-based Translation Models for Statistical Machine Translation. In *Proceedings of Human Language Technologies*, pages 394–402, Stroudsburg, PA, USA.

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

Adam David Lopez. 2008. *Machine Translation by Pattern Matching*. ProQuest.

David Mareček, Rudolf Rosa, Petra Galuščáková, and Ondřej Bojar. 2011. Two-step Translation with Grammatical Post-Processing. In *Proceedings of the 6th Workshop on Statistical Machine Translation (WMT)*, pages 426–432.

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

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

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey. ELRA.

Santanu Pal, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. 2013. A Hybrid Word Alignment Model for Phrase-Based Statistical Machine Translation. *ACL 2013*, page 94.

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.

Kristen Parton and Kathleen McKeown. 2010. MT Error Detection for Cross-lingual Question Answering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 946–954, Stroudsburg, PA, USA.

Kristen Parton, Nizar Habash, Kathleen McKeown, Gonzalo Iglesias, and Adriá 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 (EAMT)*.

Rudolf Rosa, David Mareček, and Ondřej Dušek. 2012. DEPFIX: A System for Automatic Correction of Czech MT Outputs. In *Proceedings of the 7th Workshop on Statistical Machine Translation (WMT)*, Stroudsburg, PA, USA.

Johann Roturier. 2009. Deploying Novel MT Technology to Raise the Bar for Quality: A Review of Key Advantages and Challenges. In *Proceedings of the 12th Machine Translation Summit*.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007a. Statistical Phrase-based Post-Editing. In *In Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies (NAACL-HLT)*.

Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. 2007b. Rule-Based Translation with Statistical Phrase-Based Post-Editing. In *Proceedings of the 2nd Workshop on Statistical Machine Translation (WMT)*, pages 203–206.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas (AMTA)*, pages 223–231.

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

Mihaela Vela and Josef van Genabith. 2015. Reassessing the WMT2013 Human Evaluation with Professional Translators Trainees. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT)*.

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