Building The First English-Brazilian Portuguese Corpus for Automatic Post-Editing

Felipe Almeida Costa, Thiago Castro Ferreira, Adriana Pagano, and Wagner Meira Jr
Universidade Federal de Minas Gerais
{felipealco,meira}@dcc.ufmg.br, {thiagocf05,apagano}@ufmg.br

Abstract

This paper introduces the first corpus for Automatic Post-Editing of English and a low-resource language, Brazilian Portuguese. The source English texts were extracted from the WebNLG corpus and automatically translated into Portuguese using a state-of-the-art industrial neural machine translator. Post-edits were then obtained in an experiment with native speakers of Brazilian Portuguese. To assess the quality of the corpus, we performed error analysis and computed complexity indicators measuring how difficult the APE task would be. We report preliminary results of Phrase-Based and Neural Machine Translation Models on this new corpus. Data and code publicly available in our repository.[1]

1 Introduction

Automatic post-editing (APE) is the computational task responsible for fixing systematic and repetitive errors found in black-box machine translation (MT) outputs (Vu and Haffari, 2018; Simard et al., 2007). APE is considered appealing for allowing the rapid and cheap customization of general-purpose machine translation models to specific application domains, avoiding the need for new systems to be trained from scratch (Correia and Martins, 2019). Moreover, these models deliver better machine-translated outputs to human translators, reducing human post-editing effort.

APE systems are developed based on triples aligning source sentences, machine translation outputs and human post-edits. Data is gathered through interfaces where human translators can post-edit and improve the quality of machine-translated documents (Alabau et al., 2014; Federico et al., 2014).

While human post-edits are necessary for training APE systems, there are not many publicly-available resources of this kind. For languages with pre-existing APE corpora, the scarcity of data has been solved by generating artificial triples (Junczys-Dowmunt and Grundkiewicz, 2016; Negri et al., 2018). However, even small APE corpora have not been publicly-available for most languages. For example, the WMT shared task on MT Automatic Post-Editing (APE), main source of studies in APE since 2015, provided an English-Spanish corpus in its 2015 version, an English-German APE corpus in its 2016-2018 versions (also a German-English in the 2017 one), and a novel English-Russian one in its 2019 version. In sum, only three languages have been explored in the main shared-task on APE.

In order to fulfill the gap of data scarcity in different languages, we have compiled a novel corpus for automatic post-editing from English into Brazilian Portuguese, a low-resource language. Brazilian Portuguese has fewer resources available compared to European Portuguese, an official language of the European Union and, as such, featuring in several parallel corpora. The source English texts were extracted from the WebNLG corpus (Gardent et al., 2017), which also provides a meaning representation aligned to each text. The extracted texts were automatically translated into Portuguese using a state-of-the-art neural machine translation service. Native speakers of Brazilian Portuguese post-edited the machine-translated output through a web interface specifically designed for our experiment. To assure the quality of the corpus, we submitted the machine and post-edited output to human evaluation, producing a quality label for each sentence.

[1] https://github.com/felipealco/webnlg-pt/

Proceedings of the 28th International Conference on Computational Linguistics, pages 6063–6069
Barcelona, Spain (Online), December 8-13, 2020

This paper describes our corpus compilation and reports the first results of Phrase-Based and Neural models to automatically post-edit it. The relevance of the corpus and the results produced by our baseline are discussed. Further steps in our study are presented in the concluding section of the paper.

2 Data Gathering

Source data The source English texts from our corpus were extracted from the WebNLG corpus (Gardent et al., 2017). This corpus was initially created for a Data-to-Text generation shared-task and consists of pairs matching meaning representations and their corresponding English verbalizations. Each meaning representation is a set of RDF triples extracted from DBPedia (Lehmann et al., 2015), whereas their verbalizations were collected in an experiment with human crowdworkers. Figure 1 shows an example of a set of RDF triples and their English verbalizations, each produced by a single crowdworker.

Subject	**Predicate**	**Object**
Alfred_Giles_(architect)	architect	Asher_and_Mary_Isabelle_Richardson_House
Kendall_County,_Texas	placeOfDeath	Alfred_Giles_(architect)
England	birthPlace	Alfred_Giles_(architect)

↓

Sentence 1	The architect of Asher and Mary Isabelle Richardson House Alfred Giles was born in England and died in Kendall County Texas.
Sentence 2	Alfred Giles was born in England and died in Kendall County, Texas. He designed the Asher and Mary Isabelle Richardson House .
Sentence 3	The architect Alfred Giles was born in England and he died in Kendall County, Texas. He was the architect of Asher and Mary Isabelle Richardson House.

Figure 1: Example of a single instance.

The WebNLG corpus has in total 42,901 English verbalizations for 16,095 distinct meaning representations. From this amount, we extracted a sample of 4,148 pairs, corresponding to the test partition of the original corpus.

Machine Translated Outputs Once the English texts were extracted from the WebNLG corpus, we translated them into Portuguese[2] using DeepL Translator, an increasingly popular neural MT application in industry. DeepL's superior performance over similar MT tools accounts for its selection, even though it would allow us to obtain translations into European Portuguese, and not Brazilian Portuguese, our intended target language for the experiment.

Human Post-Edits To obtain the human post-edits, we designed a web interface for this study.[3] For each instance, participants are presented with the original text, a label for its domain category and the machine translation output. They may either edit the machine translation from a text-box (so-called *free mode*), or on a *guided mode*, where a set of operations may be used to post-edit the translation. The operations were defined based on neural programmer-interpreter approaches for APE (Vu and Haffari, 2018) and consisted of *insertion to the right*, *insertion to the left*, *delete*, and *update*.

Prior to post-editing, participants received instructions requesting them (1) to transliterate entity names whenever transliteration was available, (2) not to pause the post-editing session while consulting external sources, and (3) to adapt the machine output to Brazilian Portuguese whenever necessary.

In total, we recruited a group of 37 participants to post-edit the machine translated sentences. Portuguese was their L1 and English their L2. 33 of them reported an upper-intermediate proficiency level of English while 4 reported an intermediate one. In order to prevent human errors and human biases, each machine translation output was post-edited by two independent participants.

[2] At the time this study was being conducted Brazilian Portuguese was not among the languages catered for by DeepL, only becoming available on April 2nd 2020.

[3] Publicly available in playground mode at `http://dcc.ufmg.br/~felipealco/webnlg-pt`

Human Evaluation In order to check the quality of the human post-edits, we carried out a task in which a third participant was asked to evaluate them as well as the machine-translated output on a scale ranging from *very poor*, *poor*, *medium*, *good*, to *very good*. All versions – machine translated and post-editions –, were presented on a single screen with no indication of their status, so that the participant could weigh all versions at the same time and with no prior knowledge regarding whether they were machine or human output.

3 Analysis

3.1 Post-Edit Analysis

To automatically identify types of local editions, we used a modified version of the longest common substring (LCS) algorithm to overlap machine translation output and human post-edits, highlighting the differences between both texts on character level. Figure 2 shows an example produced by our LCS adaptation.

```
Arquite[c|]to, Alfred Giles nasceu [em|na] Inglaterra.
```

Figure 2: Edition visualization. In this example 'c' was deleted and 'em' was replaced by 'na'. English gloss: An architect, Alfred Giles was born in England.

Our analysis yielded the most frequent differences between machine translations and their post-edited versions, namely `[|,]` (546), `[ó|ô]` (410), `[|"]` (395), `[|.]` (395), and `[a|o]` (384 cases). Three out of them are related to punctuation problems in the machine translation output. Spelling differences between Brazilian and European Portuguese ranked second in frequency, as is the case of words with a circumflex diacritic in the former (e.g., *Quilômetro* - Kilometer), which are spelled with an acute one in the latter (e.g., *Quilómetro*). Incorrect grammatical gender agreement between nouns and articles was the third most common post-edit, implicating editing prototypical inflections for feminine nouns *a* and for masculine ones *o*, as well as *a* and *o* as singular feminine and masculine definite articles in Brazilian Portuguese.

3.2 Sentence Length

Although sentence length plays an important role, assuming longer sentences are more difficult to translate, this did not prove to be the case in our Post-Editing experiment. TER scores and source sentence length correlated weakly with a -0.27 Pearson correlation coefficient. Size of triples set did not emerge as a measure of sentence complexity either, scoring -0.23 on Pearson's coefficient. As expected, there was strong correlation between sentence length and triple set size (0.87 Pearson's coefficient).

3.3 Complexity Indicators

Repetition Rate (RR) is a complexity indicator which measures the repetitiveness within text. Higher repetitiveness implies that correction patterns learned from the training partition are also useful to the test partition (Bojar et al., 2016; Bojar et al., 2017; Chatterjee et al., 2018). RR for source, machine translation output and post-edits was 42.33, 44.63 and 32.02, respectively. These results are considerably high in comparison to all previous editions of the WMT APE shared task. The highest score of all editions was obtained by the En-Ru corpus in 2019 with 18.25 for the source part (Chatterjee et al., 2019). A high RR such as the one obtained in our corpus allows more efficient training of machine learning algorithms for APE.

MT Quality is also an important aspect to measure difficulty of an APE task, i.e. the higher the quality of automatically translated texts, the fewer the post-edits needed, thus reducing the chances for learning patterns to feed the APE task. In this study, we assess the quality of the machine translation output by comparing it with its human post-edited version using BLEU (Papineni et al., 2002) and TER (Snover et al., 2006). Excluding low quality post-edits (labeled as *poor* or *very poor*), our corpus yielded a 81.51 BLEU and a 12.72 TER.

TER Distribution is often used in APE modeling to show the proportion of sentences that require little post-editing. Having many sentences that are close to the target or that do not require post-editing is not desirable, since APE models tend to make unnecessary editions (Chatterjee et al., 2018), which could decrease their BLEU and TER scores. In our corpus, 74.4% of the sentences (3088 out of 4148) range from 0 to 10 on TER score, out of which 74.2% (2292) were left unedited by at least one participant.

Overall Complexity The three indicators must be weighed to estimate APE complexity. Although Repetition Rate is intuitively relevant to APE, it does not play an important role when working with high quality MT (Chatterjee et al., 2019), which is the case in this corpus, as confirmed by TER Distribution. MT Quality and TER Distribution characterize this corpus as medium difficulty level for APE.

4 Automatic Post-Editing Experiment

We trained phrase-based and neural models to automatically post-edit the machine translations from the corpus. We explored the impact of using only high quality translations on each model. Thus, for each model type, we produced a model using all translations and another one using only high quality ones according to the human evaluation.

Data From the 8,296 triples in our corpus, we filtered out the ones where the machine translation output had been edited likewise by the two human post-editors, leaving 7,247 unique triples. We then selected texts evaluated as medium, good, and very good (6,646 triples). This high quality triple set was split into training, test, and development on the proportion 0.6, 0.2 and 0.2, resulting in 3,987, 1,330, and 1,329 triples, respectively. The training set was labeled GOOD training. Low quality sentences initially filtered out were added to the GOOD training set to create the ALL training set. Hence, the GOOD training, test and development sets contain high quality post-edits only, while the ALL training set contains all post-edits.

Phrase-based Model We used the *Moses* toolkit (Koehn et al., 2007) as our PB model. We extract and score phrase sentences up to the size of 18 tokens, which is the average length of sentences in the ALL training set. Besides Moses' default ordering model – distance based with distance 6 – , we used two other bidirectional ones. For the language model, we trained a 5-gram LM on the Portuguese Wikipedia dump using KenLM (Heafield et al., 2013). At decoding time, we used a 1000 stack size. We trained two models using each training set, ALL and GOOD. In the tuning phase, we used MERT (Bertoldi et al., 2009) with $k = 60$ in the development set for both models.

Neural Model We used the *Nematus* framework (Sennrich et al., 2017) to implement our Transformer encoder-decoder architecture. The model was trained using stochastic gradient descent with Adam (Kingma and Ba, 2015) ($\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$) and evaluated on the development sets after every 3,000 updates. Early stopping was applied with patience 5 based on cross-entropy. *Byte-pair encoding* (BPE) (Sennrich et al., 2016) was used to segment the tokens of source and target sides. Encoder, decoder and softmax embeddings were tied, whereas decoding was performed with beam search of size 5 to predict sequences with length up to 100 tokens. Both encoder and decoder consisted of $N = 6$ identical layers. Word embeddings and hidden units were 256D each, whereas the inner dimension of feed-forward sub-layers were 2048D. The multi-head attention sub-layers consisted of 8 heads each. A dropout of 0.1 was applied to the sums of word embeddings and positional encodings, to residual connections, to the feed-forward sub-layers and to the attention weights. At training, models had 4000 warm-up steps and label smoothing of 0.1.

Results Table 1 depicts BLEU and TER scores of DeepL machine translation (e.g., Baseline) and the two variations of each model, GOOD and ALL. The train scores in this table refer to the ALL training set. Some sentences in this set had not been seen in models trained with the GOOD training set since the GOOD training set is a subset of ALL training set, as explained in under Data in this section. For the test set, the PB model trained on the GOOD training set yielded the highest results, though not higher than the Baseline and the PB trained on the ALL training set. No remarkable improvement was seen with either of the two models, PB and neural, when using only the high quality translations.

Model	Train (ALL)		Dev		Test	
	BLEU	TER	BLEU	TER	BLEU	TER
Baseline	80.16	13.82	79.40	14.09	80.20	13.88
Moses GOOD	82.07	12.61	80.31	13.63	**80.51**	**13.67**
Moses ALL	82.07	12.58	80.30	13.72	80.43	13.7
Transformer GOOD	78.52	16.47	69.16	22.87	69.67	22.09
Transformer ALL	88.42	8.64	70.50	21.52	71.15	20.91

Table 1: BLEU and TER Results of the Baseline as well as the PB and Neural models.

5 Discussion

This study introduces the first APE corpus for English and Brazilian Portuguese, the latter being a low-resource language. Similar to Shimorina et al. (2019), our corpus is based on the WebNLG dataset (Gardent et al., 2017), originally proposed for the task of Natural Language Generation and which consists of instance pairs of meaning representations and their English verbalizations. In this arrangement, the use of meaning representations along with the original text and the machine translation in the task of APE can be investigated on our corpus.

Our analysis showed the most frequent post-edits performed by human translators in our corpus, such as inserting punctuation, fixing gender agreement and adapting European Portuguese to Brazilian Portuguese. Complexity indicators showed high repetition rate in the corpus; machine translations showed good quality as evidenced by MT metrics and also by the fact that 55.3% of them (2292 out of 4148) were not edited by at least one of the post-editors. High repetitiveness indicates higher chance of learning from the training set correction patterns (Bojar et al., 2016; Bojar et al., 2017; Chatterjee et al., 2018; Chatterjee et al., 2019), whereas good quality of machine translations represents a challenge on training an APE model to outperform the machine translation tool in our corpus.

Preliminary results for our corpus with Phrased-Based and Neural APE models show PB models outperform Neural models, which may suggest that neural models cannot generalize well with small data. This points to the importance of the training data size, supporting the generation of artificial data (Junczys-Dowmunt and Grundkiewicz, 2016; Negri et al., 2018). We can also conclude that large data is better than filtered data, especially for neural models, as the metrics in Table 1 suggest. Lastly, the relatively little improvement achieved for the APE systems tested evidences the good quality of the machine translation output which makes APE a hard task in a corpus as the one used in our study.

A further step in our study is the use of multi-source APE systems to reach better results (Chatterjee et al., 2016; Libovický et al., 2016). Specifically for this corpus, not only source sentences, but also RDF triples can be included in the encoding phase along with machine translations. In addition, using artificial data could improve APE as (Chatterjee et al., 2018) suggested, which would also fix the problem of having small data. Once our APE reaches a reasonable level of improvement, it can be used to finalize the translation of the entire WebNLG corpus into Brazilian Portuguese.

Acknowledgments

This research was partially funded by the agencies CNPq, CAPES, and FAPEMIG. In particular, the researchers were supported by CNPQ grant No. 310630/2017-7, CAPES Post doctoral grant No. 88887.508597/2020-00, and FAPEMIG grant APQ-01.461-14. This work was also supported by projects MASWeb, EUBra-BIGSEA, INCT-CYBER, and ATMOSPHERE. The authors also wish to express their gratitude to Deepl for kindly granting a license to translate our corpus, and to the students at UFMG who took part in the post-editing experiment.

References

Vicent Alabau, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Ulrich Germann, Jesús González-Rubio, Robin Hill, Philipp Koehn, Luis A Leiva, et al. 2014. Casmacat: A computer-assisted translation workbench. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 25–28.

Nicola Bertoldi, Barry Haddow, and Jean-Baptiste Fouet. 2009. Improved minimum error rate training in moses. *The Prague Bulletin of Mathematical Linguistics*, 91(2009):7–16.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, 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: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany, August. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark, September. Association for Computational Linguistics.

Rajen Chatterjee, José G. C. de Souza, Matteo Negri, and Marco Turchi. 2016. The FBK participation in the WMT 2016 automatic post-editing shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 745–750, Berlin, Germany, August. Association for Computational Linguistics.

Rajen Chatterjee, Matteo Negri, Raphael Rubino, and Marco Turchi. 2018. Findings of the WMT 2018 shared task on automatic post-editing. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 710–725, Belgium, Brussels, October. Association for Computational Linguistics.

Rajen Chatterjee, Christian Federmann, Matteo Negri, and Marco Turchi. 2019. Findings of the WMT 2019 shared task on automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 11–28, Florence, Italy, August. Association for Computational Linguistics.

Gonçalo M. Correia and André F. T. Martins. 2019. A simple and effective approach to automatic post-editing with transfer learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3050–3056, Florence, Italy, July. Association for Computational Linguistics.

Marcello Federico, Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marco Trombetti, Alessandro Cattelan, Antonio Farina, Domenico Lupinetti, Andrea Martines, Alberto Massidda, Holger Schwenk, Loïc Barrault, Frederic Blain, Philipp Koehn, Christian Buck, and Ulrich Germann. 2014. The MateCat tool. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 129–132, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada, July. Association for Computational Linguistics.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria, August. Association for Computational Linguistics.

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: Volume 2, Shared Task Papers*, pages 751–758, Berlin, Germany, August. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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 Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Jindřich Libovický, Jindřich Helcl, Marek Tlustý, Ondřej Bojar, and Pavel Pecina. 2016. CUNI system for WMT16 automatic post-editing and multimodal translation tasks. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 646–654, Berlin, Germany, August. Association for Computational Linguistics.

Matteo Negri, Marco Turchi, Rajen Chatterjee, and Nicola Bertoldi. 2018. ESCAPE: a large-scale synthetic corpus for automatic post-editing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL'16, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain, April. Association for Computational Linguistics.

Anastasia Shimorina, Elena Khasanova, and Claire Gardent. 2019. Creating a corpus for Russian data-to-text generation using neural machine translation and post-editing. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 44–49, Florence, Italy, August. Association for Computational Linguistics.

Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. 2007. 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, June. Association for Computational Linguistics.

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*, volume 200. Cambridge, MA.

Thuy-Trang Vu and Gholamreza Haffari. 2018. Automatic post-editing of machine translation: A neural programmer-interpreter approach. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3048–3053, Brussels, Belgium, October-November. Association for Computational Linguistics.