Augmented Spanish-Persian Neural Machine Translation

Benyamin Ahmadnia and Raul Aranovich
Department of Linguistics, University of California at Davis, U.S.A.
{*ahmadnia, raranovich*}*@ucdavis.edu*

Keywords: Computational Linguistics, Natural Language Processing, Machine Translation, Low-resource Language Pairs.

Abstract: Neural Machine Translation (NMT) performs training of a neural network employing an encoder-decoder architecture. However, the quality of the neural-based translations predominantly depends on the availability of a large amount of bilingual training dataset. In this paper, we explore the performance of translations predicted by attention-based NMT systems for Spanish to Persian low-resource language pairs. We analyze the errors of NMT systems that occur in the Persian language and provide an in-depth comparison of the performance of the system based on variations in sentence length and size of the training dataset. We evaluate our translation results using BLEU and human evaluation measures based on the adequacy, fluency, and overall rating.

1 INTRODUCTION

Complexity of the Statistical Machine Translation (SMT) (Koehn et al., 2003) paradigm due to training of multiple components and inability to capture long-term dependencies has diverted the attention of researchers to neural-based approaches. With the evolution of Neural Machine Translation (NMT) (Bahdanau et al., 2015) as a modern methodology for translation, MT systems have proved to exhibit further improvements over the SMT systems. The encoder and decoder form the central units of the NMT systems, providing an end-to-end system for translation from a source natural language to the target one (Cho et al., 2014). The encoder performs transformation of the source sentences of variable length into fixed length vectors, and the decoder then, generates a variable-length output from the fixed representations.

Basically, NMT systems were developed based on sequence to sequence learning on languages utilizing a gated Recurrent Neural Network (RNN) which comprises of Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) encoder-decoder architecture. Eventually, improvements on the NMT model was achieved using a LSTM with "global" and "local" attention-based mechanisms for both the encoder and decoder (Luong et al., 2015). Global attention-based LSTM considers all the source positions at each timestamp whereas local attention attends only to a subset of the source position. However, much of the positive outcome of NMT systems owe to the availability of large bilingual training datasets. Consequently, in the case of low-resource languages, there is a concern on the performance of NMT systems being particularly lower as compared to languages with largely available parallel corpus.

In this paper, we investigate the effectiveness of NMT systems on Spanish-Persian low-resource translation. Our motivation for choosing Spanish and Persian as the case-study is the linguistic differences between these languages, which are from different language families and have significant differences in their properties, may pose a challenge for MT.

Low-resource languages, also known as resource poor, are those that have fewer technologies and datasets relative to some measure of their international importance. In simple words, the languages for which parallel training data is extremely sparse, requiring recourse to techniques that are complementary to standard MT approaches. The biggest issue with low-resource languages is the extreme difficulty of obtaining sufficient resources.

Natural Language Processing (NLP) methods that have been created for analysis of low-resource languages are likely to encounter similar issues to those faced by documentary and descriptive linguists whose primary endeavor is the study of minority languages. Lessons learned from such studies are highly informative to NLP researchers who seek to overcome analogous challenges in the computational processing of these types of languages (Ahmadnia and Dorr, 2019).

Availability of Spanish-Persian parallel data in

Ahmadnia, B. and Aranovich, R.
Augmented Spanish-Persian Neural Machine Translation.
In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021) - Volume 1*, pages 482-488
ISBN: 978-989-758-484-8 ISSN: 2184-433X

digital form is limited and inadequate for performing data-driven translations. Most of the data available online are not suitable for direct usage in preparation of the corpus. The data first needs to be cleaned and carefully preprocessed for use in research which is both time-consuming and tedious. To address this issue, our goal is to prepare a high-quality bilingual corpora by collecting data covering various domains from various online and offline sources for Spanish-to-Persian translation.

We evaluate the NMT system predicted translations employing BLEU automatic measure and human evaluation measures; 1) adequacy, 2) fluency, and 3) overall rating. Additionally, we perform a comprehensive analysis on the errors in the predicted translation based on best, average and worst performance of test sentences. Performance of translated results of NMT system have been evaluated from different aspects, based on the variations in sentence length and size of training data.

This paper is organized as follows; Section 2, reviews our caste-study language issues. Section 3 describes the relevant researches related to MT, particularly in low-resource conditions. Section 4 details the architecture of NMT system. Section 5 outlines the experimental design and corpus description. Section 6 describes analyzes the results. Section 7 concludes the paper.

2 PERSIAN VS. SPANISH

Persian significantly suffers from shortage of digitally available parallel and monolingual texts. It is morphologically rich, with many characteristics shared by Urdu and Arabic. It makes no use of articles (*a*, *an*, *the*) and no distinction between capital and lowercase letters. Symbols and abbreviations are rarely used. As a consequence of being written in the Arabic script, Persian uses a set of diacritic marks to indicate vowels, which are generally omitted except in infant writing or in texts for those who are learning the language. Sentence structure is also different from that of English. Persian places parts of speech such as nouns, adverbs, and verbs in different locations in the sentence, and sometimes even omits them altogether. Some Persian words have many different accepted spellings, and it is not uncommon for translators to invent new words. This can result in OOV words.

Spanish utilizes the Latin alphabet, with a few special letters, vowels with an acute accent *(á, ú, é, ó, í)*, *u* with an umlaut (*ü*), and an *n* with a tilde (*ñ*). Due to a number of reforms, the Spanish spelling system is almost perfectly phonemic and, therefore, easier to learn than the majority of languages. Spanish is pronounced phonetically, but includes the trilled *r* which is somewhat complex to reproduce. In the Spanish IPA, the letters *b* and *v* correspond to the same symbol *b* and the distinction only exists in regional dialects. The letter *h* is silent except in conjunction with *c*, *ch*, which changes the sound into *tf*. Spanish language punctuation is very close to English. There are a few significant differences; For example, in Spanish, exclamative and interrogative sentences are preceded by inverted question and exclamation marks. Also, in a Spanish conversation, a change in speakers is indicated by a dash, while in English, each speaker's remark is placed in separate paragraphs. Formal and informal translations address several different characteristics. Inflection, declination and grammatical gender are important features of Spanish language.

A number of divergences (Dorr, 1994; Dorr et al., 2002) between low-resource (e.g., Persian) and high-resource (e.g., Spanish) languages pose many challenges in translation. In Persian, the modifier precedes the word it modifies, and in Spanish the modifier follows the head word (although it may precede the head word under certain conditions). In Persian, sentences follow a "Subject", "Object", "Verb" (SOV) order, and in Spanish, the sentences follow the "Subject", "Verb", "Object" (SVO) order (Ahmadnia et al., 2017). Such distinctions are exceedingly prevalent and thus pose many challenges for machine translation.

3 RELATED WORK

In the case of low-resource language settings, NMT models have already been explored to perform comparatively poorer than statistical models owing to the large parameter space of neural models and are often prone to overfitting.

Employing transfer learning, performance of a low-resource language pair has been improved by transferring the learned parameters from a set of high-resource languages to a set of low-resource languages. A comparison of English-Basque translation has been carried out using Google translate, NMT through OpenNMT (Kelin et al., 2017) and SMT through Moses (Koehn et al., 2007). Both NMT and SMT outperformed Google Translate when instances of training and testing sets were used from the same corpora. However, Google Translate performed better on longer, more complex sequences (Unanue et al., 2018).

One of the common approaches for dealing with

low-resource language pairs is employing monolingual data to enhance the translation prediction. According to the supervised learning method, limited parallel sentences are used along with the monolingual data for translation (Gülçehre et al., 2015; He et al., 2016; Gu et al., 2018a; Ahmadnia and Dorr, 2019).

Using unsupervised learning (Lample et al., 2018), variants of NMT and SMT models has been proposed using three principles of initializing parameters, language models and back-translation to generate parallel text. However, the semi-supervised learning method outperforms the other baseline approaches. According to Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017), low-resource languages using the universal lexical representation technique makes effective use of the availability of high-resource languages pairs indeed (Gu et al., 2018b).

NMT models are proposed by jointly incorporating RNN and Convolutional Neural Network (CNN) in order to handle longer sequences more efficiently based on the time-step and mini-batch dimensions. In the absence of a parallel corpora or for zero-resource MT, various approaches have already been proposed which are broadly classified as multilingual and pivot-based approaches (Ahmadnia et al., 2017). Using a multilingual approach, focus is on exploiting a multilingual parallel data for the translation. A multi-way, multilingual NMT has been proposed for zero-resource language where using a many-to-one strategy was found to perform better than a one-to-one strategy for zero-resource translation (Firat et al., 2016).

An NMT system for translation of multilingual languages has been proposed which uses the same single NMT architecture model (Johnson et al., 2017). It introduces the use of a token that indicates the target language at the beginning of each input source that enables it to perform multilingual translation. Another approach uses a pivot or a third language to help with the translation in the absence of parallel corpora (Chen et al., 2017). These approaches achieve zero resource translation effectively, however, difficulty lies in universal representation for multiple languages as well as increased complexity of the NMT models.

4 TRANSLATION SYSTEM ARCHITECTURE

In this paper, we used a global attention-based bidirectional LSTM on top of OpenNMT[1] as an open-source library that implements the sequence-to-sequence model with attention mechanism and performed preprocessing, training, and translation on the test dataset.

OpenNMT employs sequence-to-sequence models and supports attention mechanism. This open-source library includes vanilla NMT models as well as attention mechanism, gating, input feeding, regularization, beam-search, etc. It also supports various modules such as encoder, decoder, embedding layer, embeddings, etc.

Our model employs an encoder with a stacked 2-layer RNN with LSTM consisting of 500 hidden units for training. The encoder reads all the source individual word until it comes across the end-of-sentence symbol (<eos>) and transforms the variable-length input to a fixed vector representation known as "1-hot" encoded vector. The length of such a vector equals to the size of the vocabulary which represents a vector of zeros where only the bit position of the source word in the vocabulary is set to 1. Such a representation takes up a lot of memory and is inefficient if the size of the vocabulary is large. In order to decrease the size of the vector and to capture words with similar context, embedding is employed.

Similar to encoder, the decoder is a 2-layer RNN with LSTM consisting of 500 hidden units. Attending to all source hidden states of the input sequence, global attention is used at each time-step. An alignment vector (a_t) is computed by estimating each source hidden vector (h'_s) with the target hidden vector (h_t):

$$a_t = \frac{exp\ (score\ (h_t, h'_s))}{\sum_{s'} exp\ (score\ (h_t, h'_{s'}))} \tag{1}$$

The size of (a_t) is equal to the number of source sequences. The score function is calculated as follows:

$$score(h_t, h'_s) = h_t W_a h'_s \tag{2}$$

The (a_t) along with (h'_s) is used to derive context vector (C_t) by taking a weighted average on all the source side hidden states. A concatenation of (C_t) and (h_t) then gives the attentional hidden vector (h'_t) as follows:

$$h'_t = tanh(W_c[C_t, h_t]) \tag{3}$$

A softmax layer is finally applied to vector (h'_t) in order to produce the translated sentence in the target language.

[1] http://opennmt.net/

Generally, translation is carried out using a beam search to figure out the most adequate translation from a set of all possible candidate translations. At each level, the beam search provides a predetermined number of most probable candidate results specified by a beam width parameter which is set to a smaller value for our experiment. A larger value of beam width will be more likely to produce high-quality output sequences but at the cost of translation time and search accuracy. A beam width of size 1 would be similar to a greedy approach that selects the most likely target word. The beam search also supports a number of normalization techniques such as length normalization, coverage normalization, and end of sentence normalization.

5 EXPERIMENTAL FRAMEWORK

For the Spanish-Persian translation, we created a bilingual corpus from various sources. The corpus consists of parallel pairs of source and target sentences that span multiple domains (GNOME[2], Tanzil[3], OpenSubtitles2018[4]). Table 1 shows the data statistics:

To prepare the training corpus, we performed space-based tokenization on the text for separating the tokens. Our dataset was separated into source and target training and validation sets. The validation set helps in the selection of models during training. The preprocessing step builds dictionaries for mapping the source and target vocabularies to their corresponding indices. It then performs shuffling of the input data so that each batch contains sentences from different parts of the corpora. Sorting carried out so that sentences of similar lengths are grouped together. Sentences that are longer than a certain threshhold (set to 80) are discarded during the preprocessing stage.

A 2-layer (bidirectional) LSTM model has been trained on our parallel corpora. The trained models obtained for a total of 100 epochs[5]. The model generated at each epoch is used to translate our test data and BLEU (Papineni et al., 2001) score is computed for each translated output. This iterative process allows us to find the most optimal epoch model (highest BLEU score) among all 100 epoch models and to analyze each model for its effectiveness in translation[6]. As an optimization method, we used the Stochastic Gradient Descent (SGD) and learning rate set to 1. The initial learning rate has been decayed by a factor of 0.5 if no decrease in the validation perplexity occurred after the ninth epoch. The learning rate was decayed at each epoch thereafter.

To assess the quality of the predicted translation, we used both automatic and human evaluation. For automatic evaluation, BLEU metric (up to 3-grams precision) was employed. On the other hand, human evaluation was performed by two experts in the target language, based on the adequacy, fluency, and overall rating of each predicted output.

To further analyze of the translation results on the basis of the length of the sentences, we grouped sentences from the test set into three disjoint sets forming three sentence groups: 1) Test1 (sentences of word length 1-5), 2) Test2 (sentences of word length 6-10), and 3) Test3 (sentence of word length greater than 10). Each of these sentence groups consists of 80 sentences. For each sentence group, the BLEU scores computed to evaluate the performance of the NMT system based on variation in length of the source sentences.

6 RESULTS ANALYSIS

We trained the system for a total of 100 epochs and determined the BLEU scores (1-gram precision) for each epoch. Since SGD is used as an optimization method, the iterative method tends to produce an under-fitted model with smaller number of training epochs. However, the learning rate is decayed after the ninth epoch with a decay factor of 0.5 after every epoch to obtain the optimal model. We obtained the highest BLEU score (1-gram precision) of 41.53 at the 25th epoch which was considered as the optimal epoch model for our system.

We computed the BLEU scores obtained at the 25th epoch model (the best performing model) for 1-gram, 2-grams, and 3-grams precisions and average BLEU scores for the test sets (see Table 2). The NMT system obtained the highest average-BLEU score for Test2 dataset and lowest average-BLEU score for Test1 dataset. The higher BLEU score of Test3 dataset compared to Test1 can be attributed to the fact that although the former consists of complex, compound and long sentences, the sentences being held out from the train data have a similar structure to the sentences in the training corpus.

[2]http://opus.nlpl.eu/GNOME-v1.php

[3]http://opus.nlpl.eu/Tanzil-v1.php

[4]http://opus.nlpl.eu/OpenSubtitles-v2018.php

[5]An epoch is when the entire corpus passes through the neural network exactly once.

[6]The NMT system is trained on a single GPU

Table 1: Data statistics for Spanish-Persian translation.

Corpus	Training	Validation	Test1	Test2	Test3
Tanzil	50K	1K	1.5K	1.5K	1.5K
GNOME	100K	1.5K	2K	2K	2K
OpenSubtitles2018	150K	2K	2.5K	2.5K	2.5K
Total	300K	4.5K	6K	6K	6K

Table 2: Spanish-Persian translation results based on BLEU scores.

Test dataset	1-gram	2-grams	3-grams
Test1	41.53	27.54	20.79
Test2	44.39	34.08	28.81
Test3	43.77	33.25	25.68

Table 3: Spanish-Persian human evaluation; Fluency.

Test dataset	Expert1	Expert2	Average
Test1	4.32	5.54	4.93
Test2	4.37	5.59	4.98
Test3	4.24	4.72	4.48

Furthermore, we observed that the NMT system performs better on short sentences as compared to long sentences. The observation is consistent with previous observations in (Bahdanau et al., 2015; Zhang et al., 2017). This can attributed to the inability of LSTM to capture long term dependencies very well, particularly in low-resource conditions. Additionally, we computed the average length of the translated or predicted output for each sentence group. The length of the translated output remains more or less within the same group length but tend to produce shorter translations on the average.

Two human experts fluent in the target language evaluated the various test sets by rating each target or predicted translation against the reference translation on a predetermined scale of 1-5, where 5 indicates highest score and 1 being the least. The ratings were given based on three metrics;

- **Adequacy:** that shows how accurately the meaning conveyed in a reference translation is preserved in the translated output.
- **Fluency:** that indicates the well-formedness of the target translation.
- **Overall Rating:** that is given based on both of the two metrics (adequacy and fluency) where a high overall rating is assigned to target sentences having high adequacy and fluency scores.

The average of the fluency, adequacy and overall rating scores of the two experts for each test set computed and considered as the final scores. The experts evaluated target translations predicted by the 25th epoch model. Tables 3, 4, and 5 show the scores of Fluency, Adequacy and Overall Rating respectively, on the test sets.

Table 4: Spanish-Persian human evaluation; Adequacy.

Test dataset	Expert1	Expert2	Average
Test1	3.34	3.98	3.66
Test2	3.95	4.28	4.11
Test3	2.69	3.44	3.06

Table 5: Spanish-Persian human evaluation; Overall.

Test dataset	Expert1	Expert2	Average
Test1	3.54	4.09	3.81
Test2	3.78	4.21	3.99
Test3	3.21	3.59	3.4

All test sets achieve high fluency scores as opposed to the adequacy scores as NMT systems tend to produce syntactically correct translations. The average score ratings for BLEU and human evaluation has a negative correlation, with high BLEU score and low human evaluation score. The higher BLEU score can be attributed to the principle of BLEU metric. BLEU uses n-gram precision and for long sentences in Test3, intuitively, the n-gram matches are higher for long sentences as compared to short sentences in the other two test sets. However, although some lexical similarities may exist between reference and predicted translations, BLEU does not take into account the syntactic and semantic constructs of the predicted translation.

Figures 1 to 4 show some examples from the candidate translations predicted by best performing epoch model and classified them into best, average and worst performance based on the overall rating of the human evaluation results to analyze the quality of the results of translations predicted by the NMT system.

In Figure 1, the NMT system correctly translates the source sentence and the predicted output achieves an overall rating of 5 which means it is perfectly adequate and fluent. In Figure 2, NMT system translates the source sentence to “thank you for your advice”. There is a mistranslation of the word “ayuda” which is partially adequate but completely fluent translation. Although the predicted output is completely inade-

Source:	Que Dios te bendiga a ti y a tu familia
Reference:	خداوند شما و خانواده تان را برکت دهد
Predicted output:	باشد که خداوند به شما و خانواده شما برکت دهد

Figure 1: The best translation performance.

Source:	Gracias por su ayuda
Reference:	با تشکر از نصیحت شما
Predicted output:	ممنون از کمک شما

Figure 2: The worst translation performance.

Source:	Prometo que no lo volvería a hacer
Reference:	قول می دهم که دیگر این کار را نکنم
Predicted output:	قول می دهم این کار را نکنم

Figure 3: The best translation performance.

Source:	¿Cómo sueles pasar tu tiempo en casa?
Reference:	چگونه معمولًا وقت خود را در خانه می گذرانید؟
Predicted output:	معمولًا چگونه وقتتان را در خانه می گذرانید؟

Figure 4: Adequate translation performance.

quate, the NMT system still produces a fluent translation. In Figure 3, predicted output by the NMT system is perfectly fluent and adequate. The overall rating given by all the three evaluators is 5. In Figure 4, the predicted output is perfectly fluent and almost adequate as the phrase "en casa" is omitted in the translation predicted by the NMT system achieving an overall rating of 4.

According to the above example of best, average, and worst performance translations, our observations are listed as follows:

- In most cases, NMT systems tends to generate fluent and syntactically correct translations, even in instances where the predicted output is completely inadequate.
- However, the fluency of NMT systems reduces with increase in the length of the sentences and with complex and compound sentences which consists of independent or dependent clauses. This may be attributed to the inability of NMT systems to capture long-term dependencies.
- NMT systems often mistranslate named entities such as person names, location as well as numbers which reduces the preciseness and accuracy of the translation. Often, NMT systems trades off adequacy for fluency of the predicted translations.
- Although NMT systems use input feeding mechanism to keep track of past alignments, we observe some source words have been translated multiple times while other words have not been translated. This indicates that the use of input feeding for alignment decision is not sufficient to prevent over-translation or under-translation.

7 CONCLUSIONS

In this paper, we employed a global attentional NMT system to train and test data for Spanish-to-Persian language pair. The predicted translations evaluated based on BLEU scores. Furthermore, human evaluations based on adequacy and fluency were evaluated. We also analyzed the performance of predicted translation based on various experimental settings. Our analysis shows that the performance of NMT systems increases with the increase of the size of training corpus. We also observed that NMT systems often mistranslate named entities and compromise adequacy for fluency.

Although MT is unlikely to completely replace human translations, the effectiveness of these automatic translations have definitely surpassed professional human translators where quality can be compromised. Since the performance of NMT systems are largely influenced by the size of the training corpus, increasing the training corpus will undoubtedly improve the translation results. Incorporating linguistic resources such as monolingual data can also improve translation results.

ACKNOWLEDGEMENTS

We thank the reviewers for valuable feedback and discussions. This work was supported by the Department of Linguistics at UC Davis (USA).

REFERENCES

Ahmadnia, B. and Dorr, B. J. (2019). Augmenting neural machine translation through round-trip training approach. *Open Computer Science*, 9(1):268–278.

Ahmadnia, B., Serrano, J., and Haffari, G. (2017). Persian-Spanish low-resource statistical machine translation through english as pivot language. In *Proceedings of Recent Advances in Natural Language Processing*, pages 24–30.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.

Chen, Y., Liu, Y., Cheng, Y., and V. Li, V. (2017). A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1925–1935.

Cho, K., merrienboer, B. V., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*, pages 1724–1734.

Dorr, B. J. (1994). Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4):597–633.

Dorr, B. J., Pearl, L., Hwa, R., and Habash, N. (2002). Duster: A method for unraveling cross-language divergences for statistical word-level alignment. In *Proceedings of the 5th conference of the Association for Machine Translation in the Americas*.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1126–1135.

Firat, O., Sankaran, B., Al-onaizan, Y., Yarman Vural, F. T., and Cho, K. (2016). Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277.

Gu, J., Hassan, H., Devlin, J., and Li, V. (2018a). Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 344–354.

Gu, J., Wang, Y., Chen, Y., Li, V., and Cho, K. (2018b). Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631.

Gülçehre, Ç., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H., Bougares, F., Schwenk, H., and Bengio, Y. (2015). On using monolingual corpora in neural machine translation. *ArXiv*, abs/1503.03535.

He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T., and Ma, W. (2016). Dual learning for machine translation. In *Proceedings of the 30th Conference on Neural Information Processing Systems*.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kelin, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of 55th Annual Meeting of the Association for Computational Linguistics*, pages 67–72.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., and Zens, R. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the Association for Computational Linguistics*, pages 177–180.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54.

Lample, G., Ott, M., Conneau, A., Denoyer, L., and Ranzato, M. (2018). Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049.

Luong, T., Sutskever, I., Le, Q., Vinyals, O., and Zaremba, W. (2015). Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 11–19.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2001). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Unanue, I. J., Arratibel, L. G., Borzeshi, E. Z., and Piccardi, M. (2018). English-basque statistical and neural machine translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

Zhang, B., Xiong, D., Su, J., and Duan, H. (2017). A context-aware recurrent encoder for neural machine translation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 25(12):2424–2432.