

Bridging Philippine Languages with Multilingual Neural Machine Translation

Renz Iver D. Baliber¹, Charibeth K. Cheng¹, Kristine Mae M. Adlaon^{1,2} and Virgion H. Mamonong²

¹De La Salle University, Manila, Philippines

²University of the Immaculate Conception, Davao, Philippines

¹{renz_baliber,charibeth_cheng,kristine_adlaon}@dlsu.edu.ph

²{kadlaon,vmamonong}@uic.edu.ph

Abstract

The Philippines is home to more than 150 languages that is considered to be low-resourced even on its major languages. This results into a lack of pursuit in developing a translation system for the underrepresented languages. To simplify the process of developing translation system for multiple languages, and to aid in improving the translation quality of zero to low-resource languages, multilingual NMT became an active area of research. However, existing works in multilingual NMT disregards the analysis of a multilingual model on a closely related and low-resource language group in the context of pivot-based translation and zero-shot translation. In this paper, we benchmarked translation for several Philippine Languages, provided an analysis of a multilingual NMT system for morphologically rich and low-resource languages in terms of its effectiveness in translating zero-resource languages with zero-shot translations. To further evaluate the capability of the multilingual NMT model in translating unseen language pairs in training, we tested the model to translate between Tagalog and Cebuano and compared its performance with a simple NMT model that is directly trained on a parallel Tagalog and Cebuano data in which we showed that zero-shot translation outperforms a directly trained model in some instances, while utilizing English as a pivot language in translating outperform both approaches.

1 Introduction

Neural Machine Translation (NMT) is an approach that trains a single neural network to translate between two languages using a parallel dataset. In a recent survey paper of (Sennrich and Zhang, 2019) on low-resource NMT, it has been shown that most of the recent works is highly data-inefficient and its performance drops sharply on low-resource set-

tings. Several techniques for addressing problems of NMT in low-resource settings were also discussed, the most recent of which includes exploiting multilingual resources through Multilingual NMT, an extension of the traditional NMT that translates across multiple languages. Google’s approach in developing a multilingual translation system was discussed by (Johnson et al., 2017), the research was able to highlight how combining high-resource languages with low-resource languages helps in improving the latter’s translation quality as compared to a standalone model. Zero-shot translation is a surprising benefit of their work that bridges translation between unseen language pairs. On a practical setting, Multilingual NMT is a more efficient approach when dealing with multiple translation scenarios as compared to having a single model per language pair as it is easier to deploy and maintain it in production. An interesting area of study in Multilingual NMT has recently been explored by (Tan et al., 2019) in which they studied the effect of clustering similar languages in a Multilingual NMT, they measured the similarity of languages in two ways: Prior Knowledge (i.e. Geographical proximity, Language Family, Topology) and through the learned Language Embeddings. They have shown that clustering similar languages yields better translation results as compared to randomly clustered languages. One thing that was not able to investigate by (Tan et al., 2019) is the effectiveness of a many-to-many multilingual NMT and zero-shot translation on closely related languages (e.g. Philippine Language Family).

The Philippines is home to 186 languages based on (Eberhard and Gary). 184 of these are considered to be living and 2 are already extinct. 175 of the living languages are indigenous and 9 are non-indigenous. Additionally, 37 in the living are institutional, 68 are developing, 38 are vigorous, 30 are in trouble, and there are 11 dying languages.

There are also 3 unestablished languages. According to (McFarland, 2004), the Philippine languages may be grouped into language families, which includes the northern group (Ilokano, Pangasinan, and Kapampangan), and a central group (Tagalog, Bicolano, Hiligaynon, and Cebuano). (McFarland, 2004) also mentioned that the Philippines is experiencing a period of language convergence, a type of linguistic change in which languages structurally resembles or blends into one another as a result of prolonged interaction, manifested by high volume of borrowing from large languages such as English, Tagalog, and other regionally significant languages. Philippine languages also has large amount of intersections and sharing across the languages according to the work of (Regalado et al., 2018) on classifying low-resource Philippine Languages. In recent years, there has been a number of research works on developing Neural Machine Translation systems but all of them are English-centric and are focusing on Tagalog (Tacorda et al., 2017) and Cebuano (Adlaon and Marcos, 2018). Machine Translation systems exists for other Philippine Languages according to the survey paper of (Oco and Roxas, 2018) but implemented on either Statistical, Transfer-based, or Corpus-based approach. To the best of our knowledge, only the work of (Adlaon and Marcos, 2018) has developed and investigated translation between two Philippine Languages where they developed a unidirectional translation system for Cebuano to Tagalog.

2 Related Works

In recent years, neural machine translation has been extended to support multilingual translation which has been strongly motivated by the following reasons: a) Maintaining a single model for each language pair or translation direction is challenging in an industry setting (Johnson et al., 2017); and b) Exploiting multilingual data in improving low-resource translation settings (Sennrich and Zhang, 2019).

In one of the earliest works in multilingual NMT, (Dong et al., 2015) extended the traditional NMT into a multitask learning framework by implementing a shared encoder for the source language and dedicated encoders for the target languages, this resulted in a translation performance that achieves higher BLEU scores than individually trained models. On the other hand (Firat et al., 2016) mod-

ified the traditional NMT by implementing multiple encoders and decoders sharing a single attention mechanism and also shown improvements on low resource translation performance. Extending the traditional NMT systems to handle multilingual translation would typically require some major changes in the architecture as shown in the works of (Dong et al., 2015) and (Firat et al., 2016) which is something that is addressed by the work of (Johnson et al., 2017) where their implementation of multilingual NMT does not require major changes in the architecture but only requires to add a special token in the dataset to signify the translation direction. (Johnson et al., 2017) conducted the experiments in three different settings: One to Many (e.g. English \rightarrow Spanish, Portuguese), Many to One (e.g. Spanish, Portuguese \rightarrow English), and Many to Many (e.g. English \leftrightarrow Spanish, English \leftrightarrow Portuguese), the most interesting experiment is the Many to Many setting where an evidence of transfer learning occurred when the model learned to do zero-shot translation or to translate between unseen language pairs. In zero-shot translation, since the model was trained to translate between English \rightarrow Spanish and English \rightarrow Portuguese, the model can generate acceptable translations for Spanish \rightarrow Portuguese. More recently, the work of (Tan et al., 2019) implemented a transformer-based multilingual nmt based on the works of (Johnson et al., 2017), the main contribution of their work is the language clustering in building a multilingual nmt that aims to answer the question of which languages should be grouped together in order to gain an optimal result. (Tan et al., 2019) based the clustering on two methods, using prior knowledge (e.g. language family, geographical proximity) and language embedding, their work shows that clustering languages yields better results in almost all of the translation tasks as compared to randomly grouped or individually trained models. However, the work of (Tan et al., 2019) did not conduct a many to many experiment which would have been interesting given that it enables zero-shot learning and it is intriguing to see the effect of zero-shot learning on closely related language pairs.

2.1 Multilingual NMT for Philippine Languages

We based our approach in the work of (Johnson et al., 2017) that presented a simple and effective approach to multilingual NMT by prepending a

special token on the dataset and having no modification to the choice of NMT architecture.

3 Data

A parallel corpora sourced from the OPUS project (Tiedemann, 2012) consisting 6 language pairs was used to train two of our models. Table 1 shows the number of instances that were extracted from the corpus for each language pair. To evaluate the zero-shot translation capabilities of the model, we sourced a small number of test data from (Witneses, 2014) and were able to get 191 rows of parallel data per language pair in the permutations of Philippine languages to Philippine languages translation.

Language Pair	Instances
English-Bicolano	264,106
English-Cebuano	728,739
English-Hiligaynon	428,302
English-Ilocano	423,329
English-Pangasinense	70,112
English-Tagalog	1,638,906

Table 1: Number of parallel data collected from OPUS project

In an attempt to investigate on the relation or effect of clustering multiple languages to machine translation performance, we visualized the embeddings of selected instances in Table 1. We selected 10 instances for each language pair and generated the word embeddings of the tokens extracted and used t-SNE for high dimensional data projection. We used the same configuration of (Wattenberg et al., 2016) except for the number of iterations as it was observed that it had reached a point of stability at iteration 1000. Learning rate was set to 10 and we used perplexity configurations of 2, 5, 30, 50, and 100.

While perplexity is a tuneable hyper-parameter that is used to estimate the number of close neighbors a point has, Figure 1 shows that in all of the perplexity configurations, 5 observable clusters were formed. We zoomed-in to the clusters and isolated the points belonging to each group as seen in Figure 1.

It was observed that articles *the* and *a* are grouped in Clusters 1 and 4, respectively, while symbols . (period) and , (comma) in Clusters 2 and 5, respectively. While words like *classmates* and *study* were found clustered together in Cluster 3.

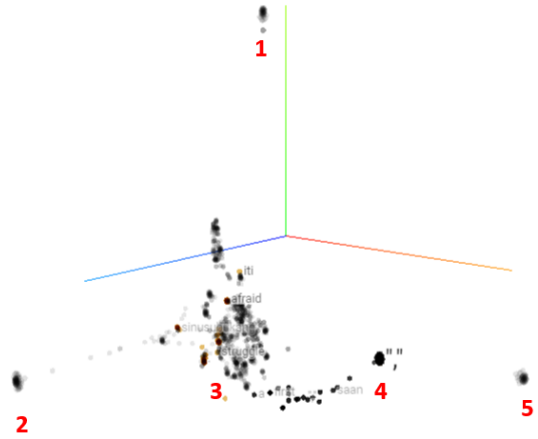


Figure 1: t-SNE Projection of the MNMT dataset.

Clustering in figure 1 shows that both semantic and syntactic relations of words in multilingual pairs could be captured but is highly evident for the English language (as shown in Figure 1) as it is the language present in all of the pairs.

With a goal of being able to aid the automatic translation of Mother-Tongue Based Multilingual Education (MTB-MLE) materials used by the Department of Education in the Philippines; a separate work was conducted gathering Tagalog-Cebuano parallel dataset in the academic domain. We continued to gather and build an academic domain-specific parallel corpus for Tagalog and Cebuano languages as initiated by the work of (Adlaon and Marcos, 2019). Due to the absence of parallel data in this domain, we manually identified specific articles and categories in Wikipedia to be extracted. Categories that were identified should be somehow related to the MTB-MLE topics. Wikipedia as the dataset source, has aligned article pairs that may range from being almost completely parallel to containing almost no parallel sentences. Table 2 shows the number of pairs that were extracted per category.

Categories	Instances
PH Regions	1537
PH Provinces	1929
PH Cities	2044
PH Tourism Attractions	1000
PH Presidents	1528
List of PH threatened species	8201

Table 2: Number of parallel data extracted from Wikipedia. PH is an abbreviation for Philippines.

3.1 Preprocessing

For the preprocessing step, we followed the steps in (Hieber et al., 2017) where we normalized punctuations, removed non-printing characters, filtered sentences longer than 100, and performed BPE (Byte pair encoding) with 50,000 joint operations. We also utilized the Moses SMT scripts from (Koehn et al., 2007) for the preprocessing steps above except for the Byte Pair Encoding step where we used the original work of (Sennrich et al., 2015). To follow the implementation of (Johnson et al., 2017) we also prepended a special token on the source data to specify the target language (i.e. <2CEB> Good Morning → Maayong Buntag)

3.1.1 Wikipedia

To extract articles from Wikipedia we used a python library called Beautiful Soup. The URL to be used for pulling data is unique per language following a format where category, topic, and language are specified. After extracting texts from the topic, we then manually aligned sentences that are translations of each other.

No subword tokenization was performed in the dataset as a preprocessing mechanism. However, we considered some rules in aligning the sentences. First, morphological markers such as affixes that indicate the tense and aspect of verbs were considered to ensure accuracy of translations. Second, in case a word in the source language has no equivalence in the target language, communicative or contextual translation was used. Third, since Tagalog and Cebuano languages mostly observe the VSO (Verb-Subject- Object) syntactic structure, we maintained the same pattern in translating from Tagalog to Cebuano. However, for cases that SVO (Subject-Verb-Object) patterns are used the same SVO structures were retained. Last, in case a certain word is too specific to a certain culture or affiliation, a more generic term was used to make the translation more inclusive.

4 Methodology

There are four fundamental parts in the methodology, the collection of parallel data from various sources, preparation of the data, model development and training, and the testing and evaluation of the models. Two models were trained and developed on the model training phase, a multilingual NMT model trained on parallel English and Philippine Languages data, and a direct NMT model

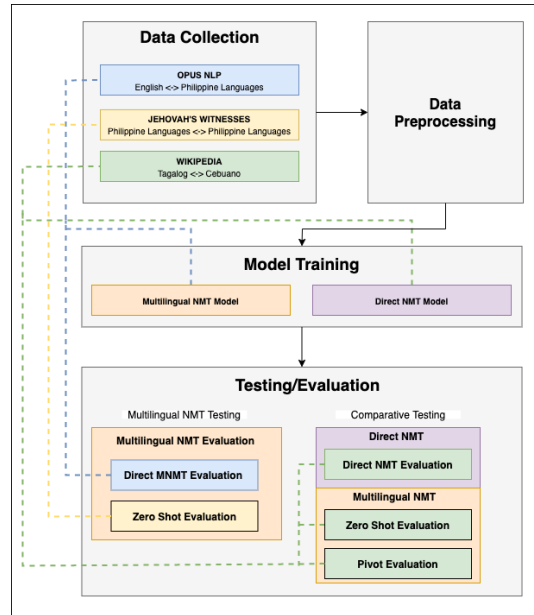


Figure 2: Different activities in the methodology

trained on a parallel Tagalog and Cebuano data. Consequently, two types of model testing was conducted, direct and comparative evaluations. The direct evaluation includes the direct evaluation of the multilingual NMT model on the English and Philippine Languages parallel test data, and evaluation of its capability to translate unseen language pairs from training through the zero shot and pivot translation using the English language on parallel data of various Philippine languages. Lastly, the comparative evaluation includes the performance comparison of a model directly trained on a Tagalog and Cebuano parallel data against a multilingual NMT model that has not seen any Tagalog and Cebuano parallel data in training, relying only on its capability to translate with zero shot translation and pivot translation.

5 Experimental Setup

All experiments were conducted using the Transformer architecture of (Vaswani et al., 2017) with relevant modifications for each model, and has been evaluated with the BLEU score based on (Papineni et al., 2002). BLEU is the averaged percentage of n-gram matches (Hwang, 2009) that for each i -gram in $i = 1, 2, \dots, N$, the percentage of the i -gram tuples in the hypothesis that also occur in the references is computed:

$$P(i) = \frac{Matched(i)}{H(i)}$$

where $H(i)$ is the number of i -gram tuples in the hypothesis. For a hypothesis of length n words, $H(1) = n$, $H(2) = n-1$, $H(3) = n-2$. Matched is computed as:

$$\text{Matched}(i) = \sum_{t_i} \min \left\{ C_h(t_i), \max_j C_{h_j}(t_i) \right\}$$

where t_i is an i -gram tuple in hypothesis h ; $C_h(t_i)$ is the number of times t_i occurs in the hypothesis; $C_{h_j}(t_i)$ is the number of times t_i occurs in reference j . BLEU is then computed as:

$$\text{BLEU}_a = \left\{ \prod_{i=1}^N P(i) \right\}^{1/N}$$

5.1 Multilingual NMT (MNMT)

We combined all the language pairs gathered from the OPUS project dataset shown in Table 1 forming a total of 7082808 instances. We then extracted 1000 instances per bilingual pair for validation and testing. There were approximately 129M words in the corpus with which 1.2M words are unique. Each instance has an average of 18 words.

JoeyNMT Framework from (Kreutzer et al., 2019) was used in training the MNMT models. Cross-entropy was used for loss computation, Adam for optimizer, and perplexity (ppl) for early stopping mechanism.

5.2 Direct NMT (DNMT)

The OpenNMT-tf toolkit (Klein et al., 2018) was used in training the model for the Direct NMT (a model trained on the gathered Wikipedia dataset). In training the model, we used a Transformer architecture with an initial learning rate of 2.0, and LazyAdam for the optimizer. No early stopping mechanism was applied. Training was stop at 20,000 steps with 7.43 and 4.69 loss for Tagalog \rightarrow Cebuano and Cebuano \rightarrow Tagalog respectively.

There were a total of 14514 cased instances gathered from the Wikipedia dataset with an overall average of 27 words per instance. The dataset was split into training, validation, and test sets with 60-20-20 percent ratio or 8706, 2904, 2904 instances per set respectively. Table 3 shows the distribution and details of the data used in training of the Direct NMT model. Despite this in-domain dataset split, only 4820 words out of 12298 in the test set is present in the training set.

Dataset	Words	Unique Words
Tagalog		
All	395324	36543
Train	244746	23995
Val	77483	11677
Test	73095	12298
Cebuano		
All	398107	36373
Train	246435	23832
Val	78276	11587
Test	73396	12095

Table 3: The Wikipedia Dataset Distribution (Train, Val, and Test)

Although the work of (Luong et al., 2015) have filtered out sentence pairs whose lengths exceed 50 words, we retained ours as it has been observed that instances with lengths greater than 50 contains enumeration of named entities such as location, dates, and names. There are a total of 734 instances with lengths greater than 50.

6 Results and Discussion

We used two approaches in multilingual translation from the MNMT model; Zero-shot and Pivot-based approach. The performance of these approaches were then compared to DNMT model which was trained on a Tagalog-Cebuano bilingual corpus. For the multilingual models, translations were detokenized before scoring. Translation results were then evaluated using SacreBleu, a python script that computes for BLEU scores except that it expects detokenized outputs (Post, 2018) from models.

Table 5 shows the scores for the dataset that was used for training the multilingual models. Since there is a lack of resource in parallel data across Philippine Languages, we leveraged on zero-shot learning to bridge translation between language pairs unseen on the training data. Table ?? shows the results of zero-shot translation from the MNMT, test data was collected from (Witnesses, 2014) which is composed of a religious journal translated across several Philippine Languages, including our required language pairs.

According to (McFarland, 2004), Hiligaynon, Bicolano, Tagalog, and Cebuano belongs to the Central Philippine languages subgroup, and at the same time, Hiligaynon is as closely related to Bicolano and Tagalog as it is with Cebuano. Our re-

	Bicolano	Cebuano	Hiligaynon	Ilocano	Pangasinense	Tagalog
Bicolano	-	13.2	16.31	10.62	9.36	3.67
Cebuano	15.22	-	16.38	10.89	9.8	3.59
Hiligaynon	12.98	11.34	-	8.71	8.22	2.58
Ilocano	8.9	10.15	3.27	-	5.84	2.91
Pangasinense	10.91	11.14	9.39	8.34	-	4.33
Tagalog	13.69	13	16.33	9.53	7.91	-

Table 4: Zero-shot translation results for Philippine Languages

sults in Table 4 supports the rationale of the Central Philippine language subgroup in which Hiligaynon almost always perform better in translating with Bicolano and Tagalog, as it is with Cebuano. Unlike the Central Philippine subgroup, the Northern languages, Ilocano and Pangasinense, are less likely to be closely related with each other as shown by (McFarland, 2004) where they have examined some grammatical structure of the Central Philippine languages and the Northern Philippine languages, and showed how strikingly different are the grammatical structures of the languages with each other, particularly in the Northern subgroup. Our results in Table 4 also shows a resemblance of this through the performance of Ilocano and Pangasinense, in which we expected this language pair to perform better with each other, considering that they belong in a single subgroup, but the results show otherwise.

Language Pair	en \rightarrow tgt	en \leftarrow tgt
English-Bicolano	51.51	54.18
English-Cebuano	46.11	44.36
English-Hiligaynon	48.38	45.37
English-Ilocano	37.27	39.13
English-Pangasinense	28.15	32.31
English-Tagalog	36.61	40.31

Table 5: Test results from the Multilingual NMT on English (en) to various target (tgt) languages

6.1 Intrinsic Evaluation

One of the benefits of Multilingual NMT is enabling zero-shot translation where the model can learn to translate between unseen language pairs during training through transfer learning, this approach is also known as implicit bridging. Prior to the emergence of the transfer learning approach, translating between unseen language pairs is possible through explicit bridging or by translating a source language to a pivot language before translat-

ing it again to the target language, assuming that the model has been trained to translate between source to pivot, and pivot to target. Both implicit (MNMT-Zero shot) and explicit (MNMT-Pivot based) approaches were implemented in these experiments and compared its performance to Direct (DNMT) approach for Tagalog and Cebuano languages as shown in Table 6.

Model	tl \rightarrow ceb	ceb \rightarrow tl
Direct NMT	9.54	10.55
MNMT-Zero shot	11.73	7.66
MNMT-Pivot based	19.08	19.19

Table 6: DNMT, MNMT-Zero shot, MNMT-Pivot based BLEU Scores using the test set in Table 3 for Tagalog(tl) and Cebuano(ceb).

Result shows that for MNMT models, Pivot-based performs better over Zero shot for tl \leftrightarrow ceb translations which is in agreement with the result of the work of (Johnson et al., 2017) where explicitly bridged model obtains higher score over implicitly bridged model for Spanish-Japanese translations.

Model	Tagalog	Cebuano
DNMT	4820	4637
MNMT	9383	8741

Table 7: Overlapping Words

While supervised models such as that of the DNMT are expected to show better performance over unsupervised models as evident in the works of (Hokamp et al., 2019), and (Ahia and Ogueji, 2020) for low-resourced languages, ours performs the least. We attribute this result to several factors. First, there is a disparity in terms of the size of the dataset that was used in training the models. Intuitively, if we look at the overlapping words found in the training vocabulary and test vocabulary of the DNMT and MNMT models in Table 7, vocabulary overlap is higher in the MNMT model than

in DNMT despite the test set being an in-domain split from DNMT. Second, the loss value when the training was stopped in the DNMT model is still considerably high compared to other NMT works where a loss value of less than 1 or when model is already able to generalize (loss value is not anymore changing over a certain number of time steps). This is evident in Table 6, where ceb→tl obtained higher BLEU score with lower loss value compared to tl→ceb.

6.2 Morpho-Syntactic Divergence

Tagalog and Cebuano verbs are conjugated by aspect (complete¹, progressive², contemplative³) or mood. In one of the translation instance the words *pagkabungkag*, *pagkaguba*, *kalaglagan* are words that synonymously mean *pagkawasak* in Tagalog. The action word */pagkawasak/* in the source text is in the complete aspect which both */pagkabungkag/*, and */pagkaguba/* exhibits because of the prefix */pagka-/*. However, *kalaglagan* is somehow a word suggesting a more figurative sense which translates to "the fall" in English. Connective words, such as *bisan pa (niana)*, *apan*, are words equivalent to *gayunpaman* in Tagalog that also express result of a certain course of action. However, in MNMT-Zero Shot, this connecting word is absent.

In the same instance, incorrect translation is manifested in DNMT where no lexical equivalent could be referred to the word in the source text *tower* where both Zero Shot and Pivot based got correctly. However the word *amateur* is seen in DNMT which belongs to the same lexical category (noun) and position as that of the source.

6.3 Lexico-Syntactic Divergence

Lexico syntactic is concerned on how the words in a sentence were put together to generate grammatically acceptable statements. In this section we describe the varying degree on how the generated translations were stringed in the four models. DNMT model generated a completely different sentence as compared to the reference sentence. On the other hand, MNMT-Zero shot model was able to translate some parts of the source sentence, however some words were expressed in English as seen in this line:

“...includes details of what happened to be

¹Indicates the action has been completed.

²Indicates the action is still being in progress.

³Indicates the action is not done but anticipated.

in the form of the Chinese fu ... into the entrance of Sprout.”

Moreover, the particle *ng* was deleted when the translation was made to the target language equivalent to *sa*:

Source: “... *pumasok sa pasukan ng gusali.*”

Google Translate: “... *misulod sa pultahan building.*”

This particle is a linker word of one word to another like in the phrase *pultahan building* should have been *pultahan sa building*. In this example, the translation of the MNMT-Pivot based model generates the closest translation.

Another case is that of zero shot, there are unrelated English and Cebuano phrases that were embedded in the translation. Phrases like “... *different from what is like...*” and “... *intindihon ug intawon...*” were seen which aren’t expected to form as in reference to the source sentence. Meanwhile, DNMT generated a structurally off translation as it completely do not coincide to the reference sentence. Pivot based model also do not project accuracy with its syntactic structure since there were deletions made in the translation. The phrase in Tagalog *arched vestibule ay kakaiba* is not translated in this sense. In fact, it only translated the **NP** *ang linya* with its counterpart in the source *Ang hanay*. The succeeding phrase *binuhing bato* is not a possible syntactic phrase for an inanimate object. In the real context, one could not take good care of rocks as if they were containing life. The word *binuhi* is only possible to living things. Hence, this phrase though syntactically correct but semantically impossible.

To examine further, the numeric modifier *napulo* to a word buttresses is not the same as that of the source sentence *labing-apat*. *Napulo* means **10** while *napulog upat* means **14**. Also, the adjectival phrase *higanteng volute* is not the same as that of the translation generated by DNMT *higanteng bulawan*. The word *bulawan* literally means gold in English is not also possible to perform an action to support a certain entity as illustrated in the sentence *higanteng bulawan nga nagasuporta*. Such phrase is not syntactically related with that from the source sentence *higanteng volute na sumusuporta*. Google Translate performs best in this specific example.

6.4 Purely-Syntactic Divergence

Tagalog and Cebuano sentences consist two general basic sentence components– predicate and topic which regularly appear in that order in basic syntactic patterns; thus it normally observes the VSO⁴; this in some cases, however, may interchange in non-basic structural patterns.

The article *ang* serves various purposes when used both in Tagalog and Cebuano. As far as the grammatical structure of these languages is concern, sentence segments that occur after the article *ang* normally point to the subject focus of the sentence. The *Ang* marks the subject focus of the sentences in the following systems DNMT, MNMT-Zero shot, and Google Translate, except for the MNMT-Pivot based in this specific example. Though noun phrases are not generally similar, as identified by the noun marker such as *orihinal gimubuhaton* (plus the additional modifier phrase – *sa Pulisya ug prominente*), *original baptism*, and *orihinal baptismuhan* in DNMT, MNMT-Zero shot, and Google Translate respectively, still it is evident that the marker provides its significant lexical function in determining the sentence focus. Regardless of differences in the specific focus of *ang*, the sentences observed syntactic divergence in the NP part.

7 Conclusion

In this paper, we presented a benchmark work on multilingual neural machine translation for major Philippine languages. We presented two translation approaches from the model that seek to address the problem of low-resource machine translation using zero-shot and pivot-based techniques. Results show that the pivot-based technique provides the closest translation compared to zero-shot and DNMT. Although zero-shot may be a good technique for MNMT, the translations it generated oftentimes contain code-switching mostly English words as discussed in the lexico-syntactic divergence section. Compared to zero-shot and pivot-based techniques, DNMT handled the retention of the sentence structure well. This could be attributed to the model’s preprocessing mechanism. We also could not see any relation to the result of our attempt to cluster languages in machine translation performance, but we see this as a future direction of study.

For future work, we plan to have a thorough extrinsic evaluation of the performance of the MNMT

⁴Verb-Subject-Object syntactic pattern

models for other PH language pairs other than Tagalog and Cebuano including further investigation of the clustering of embeddings as it relates to translation results. Also, in-depth work will be conducted on addressing the issues identified in the evaluation. In parallel to improving the MNMT models, a thorough investigation of having both morphologically-complex language pairs (e.g. Tagalog and Cebuano) versus complex and non-complex language pairs as a dataset will be studied.

References

- Kristine Mae M Adlaon and Nelson Marcos. 2018. Neural machine translation for cebuano to tagalog with subword unit translation. In *2018 International Conference on Asian Language Processing (IALP)*, pages 328–333. IEEE.
- Kristine Mae M Adlaon and Nelson Marcos. 2019. Building the language resource for a cebuano-philipino neural machine translation system. In *2019 3rd International Conference on Natural Language Processing and Information Retrieval Proceedings*, page 127–132. ACM.
- Orevaoghene Ahia and Kelechi Ogueji. 2020. Towards supervised and unsupervised neural machine translation baselines for nigerian pidgin. In *Proceedings of the AfricanNLP Workshop, International Conference on Learning Representations 2020*.
- Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732.
- David M Eberhard and F Gary. Simons, and charles d. fennig (eds.). 2020. *Ethnologue: Languages of the World*.
- Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*.
- Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *arXiv preprint arXiv:1712.05690*.
- Chris Hokamp, John Glover, and Demian Gholipour. 2019. Evaluating the supervised and zero-shot performance of multi-lingual translation models. In *Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers*, page 209–217. ACL.

- Mei-Yuh Hwang. 2009. Computing bleu scores. <http://ssli.ee.washington.edu/mhwang/pub/loan/bleu.pdf>.
- Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhipeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander Rush. 2018. Opennmt: Neural machine translation toolkit.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 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. Association for Computational Linguistics.
- Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. **Joey NMT: A minimalist NMT toolkit for novices**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 109–114, Hong Kong, China. Association for Computational Linguistics.
- Minh-Thang Luong, Hieu Pham, and Christopher Manning. 2015. Effective approaches to attention-based neural machine translation.
- Curtis D McFarland. 2004. The philippine language situation. *World Englishes*, 23(1):59–75.
- Nathaniel Oco and Rachel Edita Roxas. 2018. A survey of machine translation work in the philippines: From 1998 to 2018. In *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*, pages 30–36.
- 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.
- Matt Post. 2018. **A call for clarity in reporting BLEU scores**. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.
- Ralph Vincent Regalado, Abien Fred Agarap, Renz Iver Baliber, Arian Yambao, and Charibeth Cheng. 2018. Use of word and character n-grams for low-resourced local languages. In *2018 International Conference on Asian Language Processing (IALP)*, pages 250–254. IEEE.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.
- Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. *arXiv preprint arXiv:1905.11901*.
- Alfred John Tacorda, Marvin John Ignacio, Nathaniel Oco, and Rachel Edita Roxas. 2017. Controlling byte pair encoding for neural machine translation. In *2017 International Conference on Asian Language Processing (IALP)*, pages 168–171. IEEE.
- Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. *arXiv preprint arXiv:1908.09324*.
- Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Martin Wattenberg, Fernanda Viégas, and Ian Johnson. 2016. **How to use t-sne effectively**. *Distill*.
- Jehovah’s Witnesses. 2014. When someone you loved dies. .