Heliyon 8 (2022) e10375

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Improving neural machine translation with POS-tag features for low-resource language pairs

Zar Zar Hlaing [a], Ye Kyaw Thu [b,c], Thepchai Supnithi [b], Ponrudee Netisopakul [a,*]

[a] *Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok, 10520, Thailand*
[b] *Language and Sematic Research Technology Research Team, NECTEC, Pathum Thani, 12120, Thailand*
[c] *University of Technology Yatanarpon Cyber City, Pyin Oo Lwin, 05081, Myanmar*

ARTICLE INFO

Keywords:
Linguistic features
Part-of-speech
Universal part-of-speech
Neural machine translation
Transformer architecture

ABSTRACT

Integrating linguistic features has been widely utilized in statistical machine translation (SMT) systems, resulting in improved translation quality. However, for low-resource languages such as Thai and Myanmar, the integration of linguistic features in neural machine translation (NMT) systems has yet to be implemented. In this study, we propose transformer-based NMT models (transformer, multi-source transformer, and shared-multi-source transformer models) using linguistic features for two-way translation of Thai-to-Myanmar, Myanmar-to-English, and Thai-to-English. Linguistic features such as part-of-speech (POS) tags or universal part-of-speech (UPOS) tags are added to each word on either the source or target side, or both the source and target sides, and the proposed models are conducted. The multi-source transformer and shared-multi-source transformer models take two inputs (i.e., string data and string data with POS tags) and produce string data or string data with POS tags. A transformer model that utilizes only word vectors was used as the first baseline model for comparison with the proposed models. The second baseline model, an Edit-Based Transformer with Repositioning (EDITOR) model, was also used to compare with our proposed models in addition to the baseline transformer model. The findings of the experiments show that adding linguistic features to the transformer-based models enhances the performance of a neural machine translation in low-resource language pairs. Moreover, the best translation results were yielded using shared-multi-source transformer models with linguistic features resulting in more significant Bilingual Evaluation Understudy (BLEU) scores and character n-gram F-score (chrF) scores than the baseline transformer and EDITOR models.

1. Introduction

Natural language processing tasks are used for various purposes, one of which is machine translation (MT). Most researchers in this field have successfully built popular machine translations such as statistical machine translation (SMT) and neural machine translation (NMT). In comparison to traditional statistical phrase-based methods, NMT models have become state-of-the-art in recent years. NMT models depend on a sequence-to-sequence (seq2seq) architecture, which employs an encoder for the source sequence vector creation and a decoder for the target sequence prediction, generally with the aid of an attention mechanism [1, 2]. Seq2Seq NMT models, such as convolutional neural network models (CNNs), recurrent neural network (RNN) models, and transformer models with a self-attention mechanism [3], have recently been proposed and have achieved improved translation performance. Furthermore, numerous studies have shown that more complicated attention mechanisms [4, 5, 6, 7, 8, 9] or external syntactic features [10, 11, 12, 13, 14, 15, 16] might help NMT systems perform better. These studies have shown that NMT models benefit greatly from the incorporation of part-of-speech (POS) tags as additional syntactic information.

Annotating the source or target side, or both the source and target sides, with linguistic features such as POS or universal POS (UPOS) tags can improve the translation quality of machine translations. To the best of our knowledge, there are no studies on effectively incorporating linguistic features into NMT models for Thai–English and Thai–Myanmar low-resource language pairs. The transformer model outperforms the RNN and CNN models, particularly in translation tasks, because it allows parallelization for alleviating the problem of learning long-term memory dependencies

[*] Corresponding author.
E-mail address: ponrudee@it.kmitl.ac.th (P. Netisopakul).

https://doi.org/10.1016/j.heliyon.2022.e10375
Received 5 March 2022; Received in revised form 25 June 2022; Accepted 15 August 2022

and replaces recurrence and convolution with personal attention (self-attention) to build the dependencies between the inputs and outputs. In this study, we thus propose *transformer* models, *multi-source transformer* models, and *shared-multi-source transformer* models with additional linguistic features for low-resource language pairs. The main objective of this study is the use of additional linguistic features in NMT models to improve the translation performance of low-resource languages. As our research hypothesis, we postulate that applying POS tag information on multi-source transformer and shared-multi-source transformer architectures must be better than a standard transformer architecture using only word vectors. We assume that POS tag information will help NMT models perform better to a certain extent. POS taggers were used in our experiments to assign the correct POS tag to each word in the corpus. A POS is a grammatical category including nouns, verbs, adjectives, adverbs, etc. Many NLP tasks benefit from POS tagging, and it may help improve the quality of the machine translation.

This study presents how we applied linguistic features including POS tags to enhance NMT models. In our experiments, we used a POS tagging format such as Word|POS. To apply each translation model of the *transformer*, *multi-source transformer*, and *shared-multi-source transformer* models, POS tags are first added to the source side. Second, POS tags are applied to the target side to execute the translation models. We then incorporate POS tags to each word on both the source and target sides for each translation model. *Multi-source transformer* and *shared-multi-source transformer* models accept two inputs (i.e., string data and string data with POS tags) and produce string data or string data with POS tags as the output. The first baseline model is the transformer model that uses only word vectors. Moreover, we also use the second baseline EDITOR model to compare with our proposed models. The NMT performance on low-resource language pairs can be improved by adding linguistic information only to the source side, only to the target side, and to both the source and target sides without affecting the NMT architecture. The main contributions of this paper include the following:

- Extending the myPOS corpus and using it to develop a Myanmar POS tagger.
- Developing a Thai POS tagger for POS tagging of Thai data.
- Proving that incorporating POS tag information on the transformer architecture can improve the NMT performance for low-resource language pairs.
- Proving that incorporating POS tag information on the multi-source transformer and shared-multi-source transformer architectures can dramatically improve the NMT performance over the baseline transformer architecture.

2. Related work

Owing to the success of applying additional linguistic features in various NLP tasks, several approaches for incorporating additional linguistic information into statistical phrase-based and neural network-based systems have been proposed. However, NMT models that use linguistic features have not been studied for low-resource language pairs such as Thai and Myanmar. In our study, additional linguistic factors, such as POS tags, are incorporated into NMT models to improve the translation quality of low-resource language pairs. The following studies are related to our research.

A statistical machine translation (SMT) was the first to exploit linguistic information in machine translation. The source and target factors were implemented with the help of the Moses toolkit [17]. They were used to input or output different label sequences such as the surface forms, lemmas, POS tags, and morphological tags [18]. In SMT systems, factored models (i.e., the models that use linguistic features) were effectively developed, particularly for morphologically rich languages. Their main objective was to alleviate data sparseness issues and estimation problems. By integrating linguistic information and automatically generated word classes, a factored statistical machine translation on the language pairs of English–German, English–Czech, English–Spanish, and English–Chinese was performed [18]. It was reported that the proposed system improves the Bilingual Evaluation Understudy (BLEU) scores by up to 2% over the baseline phrase-based SMT system.

A factored statistical machine translation for two-way translation of the language pairs, Myanmar–English and Myanmar–Japanese, was proposed by Thu et al. [19]. They built machine translation models with several translation configurations using POS tags as a factor. The authors demonstrated improvements in the translation quality of Myanmar-to-English and Myanmar-to-Japanese language pairs. The authors in [20] developed a factored machine translation for the language pairs of Brazilian Portuguese and English. In their experiments, they used POS tags and morphological information as factors. Factored translation models that use additional linguistic information performed better than a baseline phrase-based statistical machine translation. When the system used the syntactic tags from the parse tree information, there was no improvement over the baseline.

Because neural network-based approaches have become increasingly prominent, the authors in [21] initially proposed a factored neural machine translation (FNMT) system. Out-of-vocabulary problems were alleviated by generating lemma and morphological tags. The incorporation of morphological information into the NMT model has been a challenging task in various studies. Additional linguistic information such as POS tags has been integrated for language models [22, 23]. In neural machine translation, incorporating additional linguistic information is beneficial [16]. The authors in [10, 11, 12, 13, 14, 15, 16] used POS tags to enhance the word representation or post-edit translation results. Incorporating the POS tags into the models causes a potential alleviation of language ambiguity and a reduction in data sparseness problems. The NMT approach with an attention mechanism was modified to produce the various linguistic features derived from the outputs [21]. Only the linguistic features of the target side were considered by the model. The authors trained the model to generate the lemma as well as the corresponding factors, such as POS tag, tense, gender, number, and person. Using priori knowledge information, the authors conducted a word-level translation with a mapping function. Their proposed factored NMT architecture increases the vocabulary coverage while decreasing the number of out-of-vocabulary words. In terms of the BLEU scores, the authors reported that their factored NMT systems outperformed the state-of-the-art NMT system.

One of the problems faced by researchers in NMT models is that the models translate every word. This does not work for words, however, particularly those in the proper noun category. The problem was solved using the POS tagging-based approach [24, 25]. Yin et al. [24] proposed the NMT model with a tag-enhanced attention mechanism, and multi-task learning was used to be jointly modeled NMT and POS tagging, where the authors applied the coarse attention mechanism on the source annotation and target hidden states for the prediction of the target POS tagging and target word. The authors in [26] developed an adaptation of the NMT model that incorporates POS tags into the attention mechanism, resulting in improved translation results. The context vector generated from source annotations and the target hidden state was utilized for target POS tagging. Then, using the context vector from the predicted target POS tags and the attention layer, the authors conducted a word prediction. They stated that they achieved better BLEU scores over the baseline models.

Perera et al. [27] proposed two approaches for incorporating POS tags into the *transformer* model for low-resource language translation from English to Sinhala. They incorporated POS information into the input embedding for the first approach and integrated POS information into the positional encoding for the second approach. The authors stated that only their first approach provides better translation performance over the *transformer* model. They proved that linguistic information, such as POS tags, indeed improves the translation quality of NMT models.

The authors in [28] proposed a multi-source neural model that utilizes two separate encoders to encode the source word sequence and linguistic feature sequences for the Turkish-to-English and Uyghur-to-Chinese translations. They applied the linguistic features only to the source side and they showed that their proposed multi-source neural model with linguistic features yields better translation quality than the baseline model utilizing word sequences. Feng et al. [29] proposed three RNN-based NMT decoding models, namely, independent decoder, gates shared decoder, and fully shared decoder for the prediction of the target word and POS tag sequences. They proved that incorporating POS tag information to the target side significantly improves the translation performance of the NMT system for Chinese-to-English and German-to-English translation pairs.

Most of the factored machine translation systems use linguistic features only on the target or source side. They do not use linguistic features on both the source and target sides. In this study, we added linguistic information such as POS tags or UPOS tags to each word on the source side, target side, and both the source and target sides. Although we desired to incorporate syntactic information into NMT models for low-resource language pairs such as Thai and Myanmar, there is no publicly available syntactic tree parser for the Myanmar language. Moreover, the authors in [20] stated that the translation model incorporating syntactic information did not perform better than the phrase-based SMT model. Thus, we integrated POS tags and UPOS tags with NMT models in our experiment. To apply the proposed models, we used the *transformer neural network architecture* without making any changes. In addition, we used multi-source translation models (i.e., *multi-source transformer* and *shared-multi-source transformer* models) according to the previous research [30]. Our research is aimed at enhancing the translation performance of low-resource languages such as Thai and Myanmar by applying additional linguistic features to NMT models. More precisely, our research is aimed at answering the following questions:

1. How much can the NMT performance increase by adding POS-tag information to transformer neural network architecture?
2. How much does it affect low-resource language pairs?

3. Materials and methods

3.1. Neural network architecture of transformer

There are two types of neural network architectures used for applying NMT models for the language pairs, including the sequence-to-sequence architecture [2] and transformer architecture [3]. In our experiment, we applied the proposed models for low-resource languages using the transformer architecture. Fig. 1 illustrates the architecture used to train the transformer model.

The transformer architecture [3] is based on the attention mechanism and a feed-forward neural network, which consists of an encoder model and a decoder model. Positional encoding, multi-head attention, and a feed-forward network are all included in these models. The encoder consists of a stack of N layers, each of which contains two sub-layers. The initial sub-layer is a multi-head self-attention structure used to acquire the attention of the input. Without a recurrent network or a convolutional network, the transformer model might not know the word-order position in a sentence. Because the model does not handle the word-order information, the authors in [3] added the positional encoding to the source word embeddings to capture the word-order position in a sentence. Self-attention is utilized in the transformer model, which has three inputs: the matrix of queries (Q), the matrix of keys (K), and the matrix of values (V). The matrix of the outputs is computed using the transformer through the equation (1):

$$Attention(H) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

The queries (Q), keys (K), and values (V) are transferred from the input H. The source representations are used to generate Q, K, and V. A position-wise fully connected feed-forward network is the second sub-layer. A residual connection [31] is utilized by the transformer, followed by layer normalization [32]. The sub-layer output is calculated in the encoder using the equation (2):

$$\begin{aligned} C^k &= LAYERNORM(ATTENTION(H^{k-1}) + H^{k-1}) \\ H^k &= LAYERNORM(FFN(C^k) + C^k), \end{aligned} \tag{2}$$

where C^k is the first sub-layer output, and H^k is the second sub-layer output in the kth layer. Although the decoder is also made up of a stack of N layers, it is not the same as the encoder. The decoder layer has an additional sub-layer to execute the attention over the encoder output.

3.2. Multi-source transformer model

In addition to the transformer model, we used the multi-source transformer model in our experiments on low-resource language pairs. The authors in [30] extended the transformer architecture [3] by adding an extra encoder and stacking an additional target-source multi-head attention component above the previous target-source multi-head attention component to enable dual-source inputs, which they effectively applied in Automatic Post-Editing. In our machine translation tasks, we also propose to utilize this multi-source (dual-source) transformer model. Fig. 2 demonstrates the architecture of multi-source transformer model. The architecture is composed of two encoders and a single decoder, one for word feature (i.e., string data {string-data}) and another for linguistic features (i.e., string data with POS tags {string-data|POS}), with the decoder producing the target sequence. Furthermore, there are two target-source multi-head attention components in the multi-source model architecture, one for each encoder. Each multi-head attention block is followed by a skip connection from the previous input and layer normalization, as in the standard transformer model. The dashed arrows in Fig. 2 indicate that sharing parameters between the two separate encoders and common embedding matrices for all encoders and the decoder. Although the two separate encoders share all parameters, they produce different activations and are combined in different parts of the decoder. The decoder in the multi-source model produces target sequence {string-data} when we utilize the word feature on the target side. By contrast, the decoder outputs the target sequence {string-data|POS} when we apply linguistic features on the target side. During training, the multi-source model is fed with a tri-parallel corpus (i.e., input {string-data}, input {string-data|POS}, and target {string-data} or {string-data|POS} in our experiment), and both input sequences are processed simultaneously during translation to produce the corrected output.

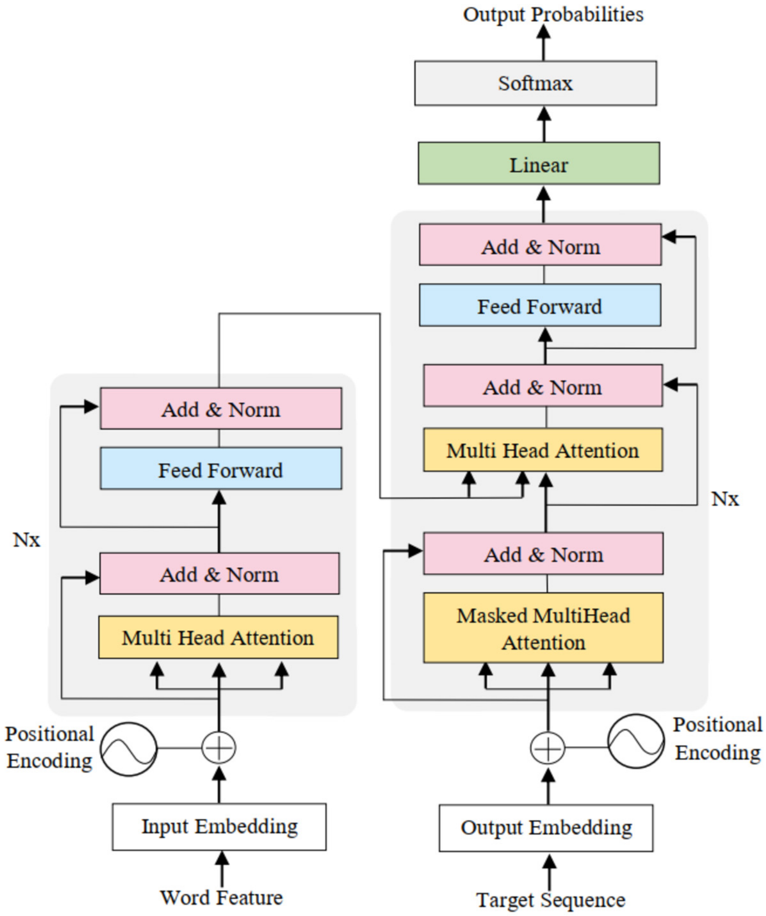


Fig. 1. Transformer architecture [3].

Table 1. Parallel dataset for Thai, English, and Myanmar.

File name	Number of sentences	Number of words	Number of unique words
train.th	20,000	139,767	8,710
valid.th	1,013	6,809	1,376
test.th	1,000	7,169	1,363
train.en	20,000	141,098	12,553
valid.en	1,013	7,245	2,113
test.en	1,000	7,176	2,201
train.my	20,000	187,547	7,712
valid.my	1,013	9,573	1,428
test.my	1,000	9,579	1,439

4. Experiment

4.1. Parallel data

Thai, English, and Myanmar sentences (without name entity tags) from the ASEAN-MT Parallel Corpus [33] are used for our experiments. This corpus is a parallel corpus in the travel domain, which is composed of six main categories, namely, people (greeting, introduction, and communication), survival (transportation, accommodation, and finance), food (food, beverage, and restaurant), fun (recreation, traveling, shopping, and nightlife), resources (number, time, and accuracy), and special needs (emergency and health). In the original ASEAN-MT corpus, we used 22,031 Thai, English, and Myanmar parallel sentences for conducting the experiments. Before being divided into training, development, and testing data, the original parallel sentences are shuffled using the *shuf* command. We divided the shuffled parallel sentences into training, development, and testing data after the shuffle process. The development and testing data cover the six main categories contained in the original parallel corpus. We used 20,000 sentences for training, 1,031 sentences for development, and 1,000 sentences for testing, respectively. The data statistics used for the experiments are shown in Table 1.

4.2. Data preprocessing

Preprocessing steps are required before conducting the training process. Owing to the absence of precise word boundaries in the Myanmar language, syllable or word segmentation is necessary to improve the machine translation quality. Although Thai and English sentences are correctly segmented in the ASEAN-MT corpus, Myanmar sentences are not. Thus, we have to segment the Myanmar data. We used the *myWord* (https://

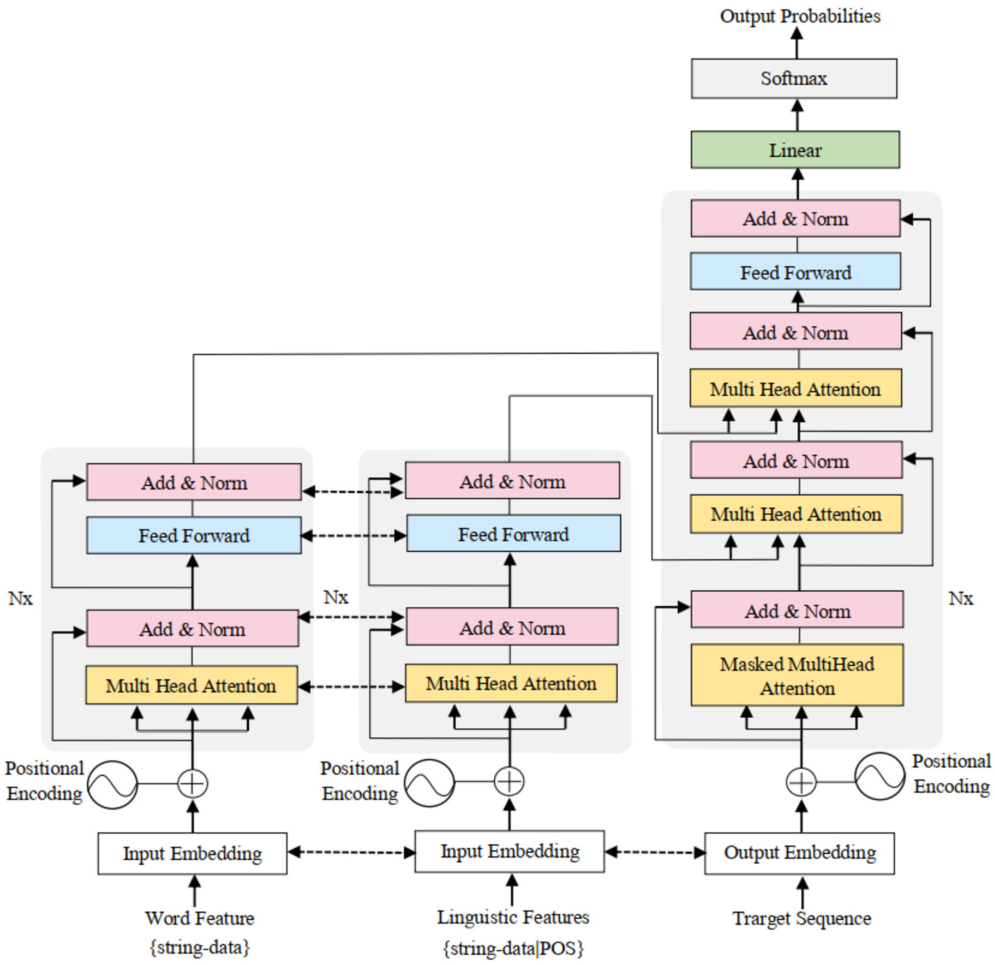


Fig. 2. Architecture of Multi-source Transformer [30]. In the case of our study, the encoder transforms string data {string-data} and the corresponding string data with POS tags {string-data|POS} separately.

github.com/ye-kyaw-thu/myWord) segmenter for Myanmar word segmentation. After conducting the word segmentation process, POS tagging was performed on each dataset of Thai, English, and Myanmar for applying NMT models using linguistic features. In the following sections, we describe the POS taggers used for each dataset.

4.3. Part-of-speech tagging

The segmented Myanmar data generated by the *myWord* segmenter tool were used for Myanmar POS tagging. The Myanmar POS tagger is required for tagging Myanmar data. Thus, we extended myPOS 2.0 created by Hlaing et al. [34] as myPOS version 3.0, and then built the Myanmar POS tagger using this extended corpus. In Section 4.3.1, we describe how to extend myPOS version 2.0 and build a Myanmar POS tagger. The English and Thai data were tagged using the Rule-based Part-of-Speech and Morphological Tagging Toolkit RDRPOSTagger [35]. In addition, we built a new Thai RDR POS tagger to compare the translation models that use the tagged data from the existing Thai POS tagger. In our experiment, we used both the existing Thai RDRPOSTagger [35] and our newly built Thai POS tagger. The translation models that utilized Thai POS data were tested using the tagged data generated from each Thai POS tagger. Section 4.3.2 presents the construction of a new Thai RDR POS tagger.

4.3.1. Extending the myPOS corpus and building Myanmar POS tagger

We used the same 15 POS tag sets [36] applied in myPOS version 2.0 to extend the myPOS corpus. These tag sets include *abb* (Abbreviation), *adj* (Adjective), *adv* (Adverb), *conj* (Conjunction), *fw* (Foreign word), *int* (Interjection), *n* (Noun), *num* (Number), *part* (Particle), *ppm* (Post-positional Marker), *pron* (Pronoun), *punc* (Punctuation), *sb* (Symbol), *tn* (Text Number), and *v* (Verb). Version 2.0 of myPOS is composed of 11,000 sentences from myPOS version 1.0 and 20,052 Myanmar sentences from our developing parallel corpora (i.e. Myanmar–Chinese and Myanmar–Korean). To extend myPOS version 2.0, 12,144 Myanmar sentences from the ASEAN MT corpus were integrated into the existing data. The extended version is regarded as myPOS version 3.0 and consists of a total of 43,196 sentences. The data statistics of the myPOS corpus extension is shown in Table 2.

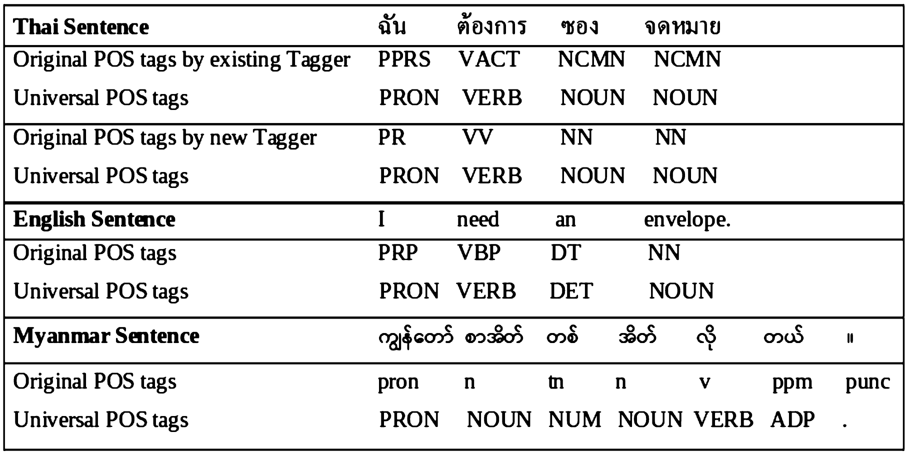
The new extended Myanmar sentences were also segmented using the *myWord* segmenter. Although the *myWord* segmenter was used to segment the words, it could not provide 100% accuracy. Thus, we manually checked and re-segmented the incorrect word segmentation before applying the POS tagging process. The revised data were then tagged using the RDR model constructed in [34]. After finishing the POS tagging process, the POS tagging errors were manually checked and cleaned. Four tagging models were trained using the extended myPOS version 3.0, including conditional random fields *(CRFs)*, hidden Markov model *(HMM)*, ripple down rules *(RDR)*, and a neural sequence labeling approach of conditional random fields *(NCRF++)*, and each model is tested using open-test data (i.e., out of training data). Table 3 describes the tagging accuracy for each model.

Table 2. Corpus information of myPOS (version 3.0).

Unit	myPOS (ver. 1.0)	Ext-1: my-zh	Ext-2: my-ko	Ext-3: ASEAN-MT my	myPOS (ver. 3.0)
Sentences	11,000	10,000	10,052	12,144	43,196
Words	239,598	103,909	106,864	114,134	564,505
Average Words/Sentence	21.78	10.17	10.64	9.40	13.07

Table 3. POS tagging accuracies (F1-score) of myPOS version 3.0 with open-test data (1,000 sentences).

Methods	Tagging Accuracies
CRFs	95.39%
HMM	94.74%
RDR	**96.53%**
NCRF + +, wordCNN-CRF-charLSTM	90.67%
NCRF + +, wordLSTM-charCNN	93.24%
NCRF + +, wordLSTM-CRF-charCNN	92.95%
NCRF + +, wordLSTM-CRF	92.08%



Thai Sentence	ฉัน	ต้องการ	ซอง	จดหมาย			
Original POS tags by existing Tagger	PPRS	VACT	NCMN	NCMN			
Universal POS tags	PRON	VERB	NOUN	NOUN			
Original POS tags by new Tagger	PR	VV	NN	NN			
Universal POS tags	PRON	VERB	NOUN	NOUN			
English Sentence	I	need	an	envelope.			
Original POS tags	PRP	VBP	DT	NN			
Universal POS tags	PRON	VERB	DET	NOUN			
Myanmar Sentence	ကျွန်တော်	စာအိတ်	တစ်	အိတ်	လို	တယ်	။
Original POS tags	pron	n	tn	n	v	ppm	punc
Universal POS tags	PRON	NOUN	NUM	NOUN	VERB	ADP	.

Fig. 3. Examples of POS and UPOS tagged sentences for Thai, English, and Myanmar.

The RDR model outperforms the other tagging models in terms of tagging accuracy. Thus, for our experiment, we used the RDR model for tagging the Myanmar data. Four POS tagging models using myPOS version 3.0 are freely available at (https://github.com/ye-kyaw-thu/myPOS).

4.3.2. Building a new Thai POS tagger

Although a Thai POS tagger (i.e., RDRPOSTagger) already exists, we built a new Thai POS tagger to compare both translation results obtained from the proposed models that utilize the existing tagger and our newly built tagger. LST20 POS tagged data [37] were used to build the new Thai POS tagger. The LST20 corpus consists of 3,164,002 words, 288,020 named entities, 248,242 clauses, and 74,180 sentences, and is annotated with 16 distinct POS tags, namely, *AJ* (Adjective), *AV* (Adverb), *AX* (Auxiliary), *CC* (Connector), *Classifier*, *FX* (Prefix), *IJ* (Interjection), *NG* (Interjection), *NN* (Noun), *NU* (Number), *PA* (Particle), *PR* (Pronoun), *PS* (Preposition), *PU* (Punctuation), *VV* (Verb), and *XX* (Others). Our newly built Thai POS tagger uses these 16 POS tags, whereas the existing RDRPOSTagger utilizes 44 POS tags [38] for Thai POS tagger and 45 POS tags [39] for English POS tagger. The different number of POS tags for each tagger used is the main difference between the two Thai POS taggers. Owing to the different numbers of POS tags, the translation results may differ according to each tagger.

Prachya et al. [37] constructed the LST20 dataset in a CoNLL2003 style format, with four columns (Word, POS tag, Named entity, and Clause boundary), each separated by a tab. There is also an empty line in the dataset that indicates the sentence boundary. Word and POS tag columns from the LST20 dataset were retrieved and converted into sentences. In the sentences, the word and POS tags were separated by the forward-slash "/." We used the converted sentence-level LST20 dataset as the training data for the construction of a new Thai POS tagger utilizing the RDRPOSTagger tagging toolkit [35]. In building the tagger, over 74K POS tagged sentences were utilized as the training data, and 1,000 sentences from the training data were taken for the close test-set (ctest) to evaluate the tagging accuracy. Our newly built Thai POS tagger provides a tagging accuracy of 96.61%. For tagging the Thai data in our experiment, we used both Thai POS taggers. The experiments for the proposed models were conducted using the Thai POS tagged data produced by each tagger. In addition to the POS tags, UPOS tags were also used for incorporation with the *transformer*, *multi-source transformer*, and *shared-multi-source transformer* models. The usage of UPOS tags will be presented in the following section.

4.3.3. Universal part-of-speech (UPOS) tags

Universal POS tags are used in the Universal Dependencies (UD) framework, which is a framework for consistent annotation of grammar (parts-of-speech, syntactic, and morphological features) in several human languages. In our experiment, linguistic features such as POS and UPOS tags were incorporated with NMT models. We used 12 UPOS tags defined by Petrov et al. [40], including *NOUN* (nouns), *VERB* (verbs), *ADJ* (adjectives), *ADV* (adverbs), *PRON* (pronouns), *DET* (determiners and articles), *ADP* (prepositions and postpositions), *NUM* (numerals), *CONJ* (conjunctions), *PRT* (particles), '*.*' (punctuation marks), and *X* (a catch-all for other categories such as abbreviations or foreign words). We replaced the original POS tags with UPOS tags using the POS-to-UPOS tags mappings for Thai, English, and Myanmar UPOS tagged data. For more details on the POS-to-UPOS tag mappings, please refer to Appendix A. Examples of POS and UPOS tagged sentences for Thai, English, and Myanmar are shown in Fig. 3.

4.4. Model configuration and evaluation

To evaluate the effectiveness of NMT models integration with linguistic features, we conducted several experiments on two-way translation of each Thai–English, Thai–Myanmar, and English–Myanmar language pair. For the implementation of our proposed NMT models, we used the *Marian* framework [41], which is written in pure C++ and has minimal dependencies. The experiments were carried out using the *transformer*, *multi-source transformer*, and *shared-multi-source transformer* models supported by *Marian* framework. Whereas the baseline transformer model uses only word vectors, the proposed transformer models integrate linguistic features into each word on either the source or target side, or both. For the *transformer* model, we conducted four different translation models, i.e., *Word-to-Word* translation model (i.e., the baseline transformer model), *Word|POS-to-Word* translation model, *Word-to-Word|POS* translation model, and *Word|POS-to-Word|POS* model for each language pair. Although the *multi-source transformer* and *shared-multi-source transformer* models are generally similar, the *shared-multi-source transformer* model shares the parameters during the training. We also experimented on two translation models for each *multi-source transformer* and *shared-multi-source transformer* model. For the first translation model, the *multi-source transformer* and *shared-multi-source transformer* models accept two inputs of string data {string-data} and POS tagged data {string-data|POS} and produce an output of string data {string-data}. For the other model, they accept those data and produce the output of POS tagged data {string-data|POS}. The first and second translation models both use the same set of inputs but produce different outputs.

During the model training, we set the parameters with a maximum sentence length of 500, a maximum bath size of 100, and a beam size of 6. For the *transformer* models, both the encoder and decoder have eight layers, with two layers for the *multi-source transformer* and *shared-multi-source transformer* models. The mini-batch size during validation is set to 64. We set 0.1 for label smoothing, 8 for multi-head attention, and 0.3 for the dropout probability. The models are trained using the *Adam* optimizer [42]. The model is validated every 5,000 updates and saved after every 5,000 iterations. Early stopping depends on the BLEU score, which is set to 10. Except for the different numbers of layers utilized in the encoder and decoder, the *transformer* models, *multi-source transformer* models, and *shared-multi-source transformer* models use the same parameter values. After the training process, the test data are translated, and statistical significance evaluations of the translated data are calculated using the *compare-mt* tool [43] with paired bootstrap resampling with 1,000 resamples.

In addition to the baseline *transformer* model, we perform further experiments for the latest transformer-based model (i.e., Edit-Based Transformer with Repositioning: *EDITOR* [44]) to use as the second baseline model for comparison with our proposed models. The EDITOR is a non-autoregressive transformer model that iteratively edits hypotheses using a novel reposition operation. In EDITOR, the deletion operation is replaced with a novel reposition operation for the lexical choice extraction from reordering decisions. Because a single reposition operation can subsume a sequence of deletions and insertions, EDITOR takes advantage of lexical constraints more effectively and efficiently than the Levenshtein Transformer [45]. EDITOR model is based on the Transformer encoder-decoder [3] and the decoder representations are extracted to make the policy predictions. The two basic operations of each refining action of the EDITOR model are reposition and insertion. We can implement the EDITOR model with *soft* lexical constraints or *hard* lexical constraints or without constraints. The authors in [46] proposed the EDITOR and Levenshtein Transformer models for two-way translation of Thai–Myanmar, Thai–English, and Myanmar–English language pairs without using any lexical constraints, and they showed that the EDITOR model achieves better translation performance in English-to-Thai, Thai-to-English, and English-to-Myanmar translation pairs. Moreover, Xu and Carpuat [44] stated that EDITOR performs more effectively than the Levenshtein Transformer when utilizing the soft lexical constraints. Thus, we conduct additional experiments on the EDITOR model with soft lexical constraints for all language pairs and use it as the second baseline model to compare with our proposed models.

For the implementation of the EDITOR model, we use the publicly available Pytorch-based FairSeq framework (https://github.com/Izecson/fairseq-editor). The same training, development, and testing data utilized in the multi-source models are used for the EDITOR model. In the experiment settings, we set dropout (0.1) and label smoothing (0.1). Every 2,000 updates, we checkpoint the models. We assign the maximum updates to 60,000 and the maximum tokens to 1,800 for model training. With an initial learning rate of 0.0005, we train the model using the *Adam* optimizer [42]. The other hyperparameters are the same as those in [45]. After the training process for the EDITOR model, we select the best checkpoint to translate the test data. Before translating the test data, we retrieve one to three words from the reference as soft lexical constraints for each source sentence, and BPE is applied to the constraint sequence. The test data are then translated using soft lexical constraints, and statistical significance evaluations are computed by the *compare-mt* tool with paired bootstrap resampling. In this study, all of the evaluation results are described in terms of the *case-insensitive Bilingual Evaluation Understudy* (BLEU) [47] and *character n-gram F-score* (chrF) scores [48].

In all result tables, we used the short notations for the proposed models. The factored models are denoted by the following notations:

- “t” indicates “translation,”
- “W” represents “words,”
- “P” indicates “part-of-speech (POS) tags,” and
- “UP” stands for “universal part-of-speech (UPOS) tags.”

For example, t(W-to-W) is the translation model for a “words” to “words” translation, t(W|P-to-W) is the translation model for a “words and POS tags” to “words” translation, t(W-to-W|P) is the translation model for a “words” to “words and POS tags” translation, and t(W|P-to-W|P) is the translation model for a “words and POS tags” to “words and POS tags” translation for the transformer models. In addition, t(W + W|P-to-W) is the multi-source translation model for a translation of two inputs (“words” + “words and POS tags”) to “words,” and t(W+ W|P-to-W|P), by contrast, is a multi-source model for translating two inputs (“words” + “words and POS tags”) to “words and POS tags.” For the translation models with UPOS tags, we used the notation “UP” instead of “P.”

5. Results and discussion

5.1. Translation results

In this section, we report the experimental results based on the BLEU and chrF scores for the proposed models. Tables 4–8 show the experiment results of the proposed models in the two-way translation of Thai–Myanmar, Myanmar–English, and Thai–English, and compare our proposed models to two baseline models that use only word vectors (i.e., *transformer* model and EDITOR model). The higher BLEU and chrF scores compared

Table 4. Experiment results for Thai-to-Myanmar translation pair. Bold numbers indicate higher BLEU and chrF scores than the two baseline models. Scores with an asterisk (*) are significantly higher than the first baseline Transformer model and scores with a dagger (†) are significantly higher than the second baseline EDITOR model at $p < 0.05$. Statistical significance is computed using the *compare-mt* tool with paired bootstrap resampling with 1,000 resamples.

Models	Thai-to-Myanmar by Existing Thai POS Tagger (44 POS tags)		Thai-to-Myanmar by New Thai POS Tagger (16 POS tags)	
	BLEU scores (%)	chrF scores (%)	BLEU scores (%)	chrF scores (%)
Baseline Transformer - t(W-W)	22.90	39.75	22.90	39.75
Baseline EDITOR *with soft constraints*	18.14	41.80	18.14	41.80
Transformer- t(W|P-to-W)	22.81†	39.70	**23.06**†	39.90
Transformer - t(W-to-W|P)	22.84†	39.42	22.84†	39.42
Transformer - t(W|P-to-W|P)	22.78†	38.41	**23.24**†	39.34
Multi-Source Transformer - t(W + W|P-to-W)	**24.81***†	40.62	**24.75***†	40.48
Multi-Source Transformer - t(W + W|P-to-W|P)	**25.34***†	41.43*	**25.03***†	41.42*
Shared-Multi-Source Transformer - t(W + W|P-to-W)	**24.17***†	40.68	**25.36***†	**42.17***
Shared-Multi-Source Transformer - t(W + W|P-to-W|P)	**24.91***†	41.43*	**24.37***†	**41.82***
Transformer- t(W|UP-to-W)	**23.21**†	40.39	**22.98**†	39.28
Transformer - t(W-to-W|UP)	22.62†	39.08	22.62†	39.08
Transformer - t(W|UP-to-W|UP)	22.61†	39.55	22.79†	38.51
Multi-Source Transformer - t(W + W|UP-to-W)	**24.64***†	40.79	**24.36***†	40.62
Multi-Source Transformer - t(W + W|UP-to-W|UP)	**24.02***†	39.76	**24.89***†	41.00*
Shared-Multi-Source Transformer - t(W + W|UP-to-W)	**24.60***†	41.77*	**25.01***†	41.35*
Shared-Multi-Source Transformer - t(W + W|UP-to-W|UP)	**25.12***†	**41.82***	**25.25***†	**42.29***

Table 5. Experiment results for Myanmar-to-Thai translation pair. Bold numbers indicate higher BLEU and chrF scores than the two baseline models. Scores with an asterisk (*) are significantly higher than the first baseline Transformer model and scores with a dagger (†) are significantly higher than the second baseline EDITOR model, at $p < 0.05$.

Models	Myanmar-to-Thai by Existing Thai POS Tagger (44 POS tags)		Myanmar-to-Thai by New Thai POS Tagger (16 POS tags)	
	BLEU scores (%)	chrF scores (%)	BLEU scores (%)	chrF scores (%)
Baseline Transformer - t(W-W)	24.92	41.73	24.92	41.73
Baseline EDITOR *with soft constraints*	14.30	42.94	14.30	42.94
Transformer- t(W|P-to-W)	24.76†	40.90	24.76†	40.90
Transformer - t(W-to-W|P)	**25.18**†	41.37	24.44†	40.84
Transformer - t(W|P-to-W|P)	24.20 †	40.72	24.66†	41.41
Multi-Source Transformer - t(W + W|P-to-W)	24.77†	**43.79***	24.77†	**43.79***
Multi-Source Transformer - t(W + W|P-to-W|P)	24.02†	**43.49***	24.56†	**43.32***
Shared-Multi-Source Transformer - t(W + W|P-to-W)	24.79†	**44.41***	24.79†	**44.41***
Shared-Multi-Source Transformer - t(W + W|P-to-W|P)	23.73†	42.59	23.97†	42.87
Transformer- t(W|UP-to-W)	24.76†	41.28	24.76†	41.28
Transformer - t(W-to-W|UP)	21.81†	40.61	24.26†	40.76
Transformer - t(W|UP-to-W|UP)	24.97†	41.50	24.19†	40.46
Multi-Source Transformer - t(W + W|UP-to-W)	24.92†	**43.25***	24.92†	**43.25***
Multi-Source Transformer - t(W + W|UP-to-W|UP)	24.71†	**43.11**	24.51†	**43.16**
Shared-Multi-Source Transformer - t(W + W|UP-to-W)	24.84†	**43.86***	24.84†	**43.86***
Shared-Multi-Source Transformer - t(W + W|UP-to-W|UP)	24.27†	**43.13**	**25.02**†	**43.85***

with the two baseline models are highlighted as bold numbers. In addition, significantly higher scores over the first baseline *transformer* model are marked with an asterisk (*), and significantly higher scores over the second baseline EDITOR model are indicated by a dagger (†).

Tables 4 and 5 describe the experiment results for two-way translation of the Thai–Myanmar translation pair that utilized both the existing Thai POS tagger using 44 POS tags and the new Thai POS tagger (i.e., our newly built Thai POS tagger using 16 POS tags). These two distinct Thai POS taggers were used for our Thai POS tagged data. There are some reasons why we had to build the new Thai POS tagger. The existing RDRPOSTagger uses 44 POS tags for the Thai data, whereas the Myanmar POS tagger only uses 15 tags. The existing Thai POS tagger uses approximately three times as many POS tags as used in the Myanmar POS tagger. Thus, for Thai-to-Myanmar and Myanmar-to-Thai translation pairs, the models may have difficulty learning to map the different POS tags during training. Moreover, we hoped to compare the translation results produced by each proposed model using two different Thai POS taggers. Hence, we built a new Thai POS tagger that only contains 16 tags using the LST20 corpus [37] and also used both the new tagger and existing RDRPOSTagger for our Thai data.

For the Thai-to-Myanmar translation pair based on the existing Thai POS tagger in Table 4, our proposed *multi-source transformer* models, i.e., t(W + W|P-to-W), t(W+ W|P-to-W|P), and t(W+ W|UP-to-W), t(W+ W|UP-to-W|UP), and the *shared-multi-source transformer* models, i.e., t(W+ W|P-to-W), t(W+ W|P-to-W|P), and t(W+ W|UP-to-W), t(W+ W|UP-to-W|UP), provide significantly higher BLEU scores than the baseline *transformer* and EDITOR models. These proposed models also provide higher chrF scores than the two baseline models. Moreover, compared to the highest score, the *multi-source transformer* model, i.e., t(W+ W|P-to-W|P), yields better BLEU scores (+ 2.44) and chrF scores (+ 1.68) than the first baseline *transformer* model and provides higher BLEU scores (+ 7.20) than the second baseline EDITOR model, and the *shared-multi-source transformer* model with UPOS tags, i.e., t(W+ W|UP-to-W|UP) produces higher chrF scores over the two baseline models. According to the new Thai POS tagger, the *shared-multi-source transformer* model with POS tags, i.e., t(W+ W|P-to-W) yields significantly higher BLEU scores (+ 2.46) and chrF scores (+ 2.42) than the first baseline *transformer* model and produces higher significance BLEU scores (+ 7.22) and better chrF scores (+ 0.37) than the second baseline model.

In Table 5, the Myanmar-to-Thai translation performance of the baseline EDITOR model (i.e., BLEU scores) decreases dramatically compared to all our proposed models. Because the syntax structures of Myanmar and Thai languages are very distinct (i.e., SOV - Subject, Object followed by Verb and SVO - Subject, Verb followed by Object), reordering is more difficult for those language translations. However, in terms of BLEU and

Table 6. Experiment results for Myanmar-to-English and English-to-Myanmar translation pairs. Bold numbers indicate higher BLEU and chrF scores than the two baseline models. Scores with an asterisk (*) are significantly higher than the first baseline Transformer model and scores with a dagger (†) are significantly higher than the second baseline EDITOR model, at $p < 0.05$.

Models	Myanmar-to-English		English-to-Myanmar	
	BLEU scores (%)	chrF scores (%)	BLEU scores (%)	chrF scores (%)
Baseline Transformer - t(W-W)	27.76	44.30	27.16	44.12
Baseline EDITOR *with soft constraints*	17.14	43.15	24.33	46.71
Transformer- t(W|P-to-W)	**27.88**†	44.24	27.05†	43.91
Transformer - t(W-to-W|P)	27.14†	43.24	**27.21**†	43.86
Transformer - t(W|P-to-W|P)	27.58†	44.13	27.00†	43.64
Multi-Source Transformer - t(W + W|P-to-W)	**28.76**†	**46.61***†	**31.72***†	**48.03***
Multi-Source Transformer - t(W + W|P-to-W|P)	**28.86**†	**46.64***†	**31.53***†	**48.03***
Shared-Multi-Source Transformer - t(W + W|P-to-W)	**28.93**†	**47.08***†	**31.24***†	**47.62***
Shared-Multi-Source Transformer - t(W + W|P-to-W|P)	**29.01**†	**46.81***†	**31.26***†	**47.66***
Transformer- t(W|UP-to-W)	27.29†	43.77	**27.38**†	44.33
Transformer - t(W-to-W|UP)	27.32†	44.12	**27.31**†	44.19
Transformer - t(W|UP-to-W|UP)	27.37†	44.03	26.64†	43.35
Multi-Source Transformer - t(W + W|UP-to-W)	**28.86**†	**46.72***†	**31.66***†	**48.30***
Multi-Source Transformer - t(W + W|UP-to-W|UP)	**28.39**†	**45.93***†	**31.67***†	**47.98***
Shared-Multi-Source Transformer - t(W + W|UP-to-W)	**29.84***†	**48.08***†	**31.73***†	**48.53***
Shared-Multi-Source Transformer - t(W + W|UP-to-W|UP)	**29.31**†	**47.08***†	**31.26***†	**47.74***

chrF scores, our proposed models achieve the best translation results. In this table, the *transformer* model that used POS tags from the existing Thai POS tagger {t(W-to-W|P)} achieves higher BLEU scores (+0.26) than the first baseline *transformer* model and provides significantly higher BLEU scores (+10.88) over the second baseline EDITOR model. The *shared-multi-source transformer* model with UPOS tags from the new Thai POS tagger {t(W+ W|UP-to-W|UP)} achieves better BLEU and chrF scores than the two baseline models. Furthermore, our proposed *shared-multi-source transformer* models with POS tags {t(W+ W|P-to-W)} yield higher chrF scores (+2.68) over the first baseline *transformer* model and (+1.07) over the second baseline EDITOR model.

According to the existing Thai POS tagger and our newly built Thai POS tagger, our proposed models provide comparable BLEU scores to the first baseline *transformer* model, but our proposed *multi-source transformer* and *shared-multi-source transformer* models provide significant chrF scores than the first baseline and those models also achieve significantly higher BLEU scores and better chrF scores over the second baseline EDITOR model. When comparing the BLEU score to the chrF score, the former is based on tokenization and is computed using tokenized word or syllable level n-grams, whereas the chrF score is not dependent on tokenization and is computed using character-level n-grams. While our proposed models could not yield higher significance BLEU scores in Myanmar-to-Thai translation, they did provide significantly higher chrF scores over the first baseline *transformer* and also achieve higher significance BLEU scores and better chrF scores over the second baseline EDITOR model. Thus, we can assume that our proposed models outperform the two baseline models in Myanmar-to-Thai translation.

For the Thai-to-Myanmar and Myanmar-to-Thai translation pairs, the *transformer*, *multi-source transformer* and *shared-multi-source transformer* models that use POS tags and UPOS tags perform better than the *transformer* model that uses only word vectors (i.e., the first baseline model). Those models also yield significantly higher BLEU scores over the second baseline EDITOR model in both translations and our proposed *shared-multi-source transformer* models with POS tags {t(W + W|P-to-W)} achieve better chrF scores over the baseline EDITOR model in Myanmar-to-Thai translation. The use of POS tags helps the decoder select the correct target words. Because Thai and Myanmar POS tags differ, the decoder has to completely learn how to map these different tags. The universal POS (UPOS) tags, by contrast, are identical in both the source and target languages and include fewer tags than the original POS tags. With the aid of POS tags, the outputs can be generated more accurately. In particular, when using the UPOS tags, the output data are better than those produced using POS tags. In general, for the Thai-to-Myanmar and Myanmar-to-Thai translation pairs, the experiment results of our proposed models that integrate linguistic features perform significantly better than the first baseline *transformer* model in terms of the chrF scores, and such models significantly outperform the second baseline EDITOR model in terms of BLEU scores.

The experiment results for the Myanmar-to-English and English-to-Myanmar translation pairs are shown in Table 6. In Myanmar-to-English translation, the *shared-multi-source transformer* model with UPOS tags {t(W + W|UP-to-W)} performs significantly better than the first baseline *transformer* model in terms of BLEU scores (+2.08) and chrF scores (+3.78). In addition, this *shared-multi-source transformer* model also yields significantly higher BLEU scores (+12.70) and chrF scores (+4.93) over the second baseline EDITOR model. Although the baseline *transformer* and EDITOR models underperform in the distinct language translation from Myanmar (i.e., SOV) to English (i.e., SVO), our proposed multi-source models achieve significantly higher BLEU and chrF scores over the two baseline models. In the reverse direction (i.e., English-to-Myanmar translation), our proposed *multi-source transformer* and *shared-multi-source transformer* models produce significantly higher BLEU and chrF scores than the first baseline *transformer* model, and those models yield higher significance BLEU scores over the second baseline EDITOR model. The *shared-multi-source transformer* model with UPOS tags {t(W+ W|UP-to-W)} provides the significant differences in terms of BLEU scores (+4.57) and chrF scores (+4.41) over the first baseline *transformer* model and significantly higher BLEU scores (+7.40) and better chrF scores (+1.82) over the second baseline. In general, the proposed models that incorporate linguistic information such as UPOS tags significantly outperform the two baseline models in terms of BLEU scores. Because those models use the same UPOS tags for both language pairs, making it easier for the models to learn the mapping and predict the target words more accurately and cause the translation models to be enhanced.

Tables 7 and 8 show the experiment results for Thai-to-English and English-to-Thai translation pairs. For Thai-to-English translation pair, the *shared-multi-source transformer* model with UPOS tags by existing Thai POS tagger {t(W + W|UP-to-W|UP)} provides better BLEU scores (+4.69) and chrF scores (+4.32) than the first baseline *transformer* model. In addition, the *shared-multi-source transformer* model with POS tags from the new POS tagger {t(W+ W|P-to-W)} provides higher BLEU scores (+4.57) and chrF scores (+5.05) over the first baseline model and better BLEU scores (+11.73) and chrF scores (+4.22) over the second baseline EDITOR model. We notice that this *shared-multi-source transformer* model performs significantly better than the two baseline models in terms of BLEU and chrF scores. Table 8 shows that the *shared-multi-source transformer* models with UPOS tags {t(W+ W|UP-to-W)} for the English-to-Thai translation pair yield significantly better BLEU scores (+5.99) and chrF scores (+5.25) over the first baseline *transformer* model and higher significance BLEU scores (+12.43) and chrF scores (+2.78) over the second baseline EDITOR model.

Table 7. Experiment results for Thai-to-English translation pair. Bold numbers indicate higher BLEU and chrF scores than the two baseline models. Scores with an asterisk (*) are significantly higher than the first baseline Transformer model and scores with a dagger (†) are significantly higher than the second baseline EDITOR model, at $p < 0.05$.

Models	Thai-to-English by Existing Thai POS Tagger (44 POS tags)		Thai-to-English by New Thai POS Tagger (16 POS tags)	
	BLEU scores (%)	chrF scores (%)	BLEU scores (%)	chrF scores (%)
Baseline Transformer - t(W-W)	30.94	45.64	30.94	45.64
Baseline EDITOR *with soft constraints*	23.78	46.47	23.78	46.47
Transformer- t(W|P-to-W)	**31.00**†	45.90	**31.14**†	46.60
Transformer - t(W-to-W|P)	**31.23**†	45.79	**31.23**†	45.79
Transformer - t(W|P-to-W|P)	30.58†	45.43	**31.09**†	46.18
Multi-Source Transformer - t(W + W|P-to-W)	**34.89***†	**49.56***†	**34.83***†	**49.33***†
Multi-Source Transformer - t(W + W|P-to-W|P)	**33.80***†	**48.58***†	**34.26***†	**49.78***†
Shared-Multi-Source Transformer - t(W + W|P-to-W)	**34.22***†	**49.70***†	**35.51***†	**50.69***†
Shared-Multi-Source Transformer - t(W + W|P-to-W|P)	**34.25***†	**49.81***†	**34.71***†	**49.57***†
Transformer- t(W|UP-to-W)	30.61†	45.73	30.70†	46.12
Transformer - t(W-to-W|UP)	**31.02**†	45.58	**31.02**†	45.58
Transformer - t(W|UP-to-W|UP)	**31.18**†	45.53	27.52†	43.26
Multi-Source Transformer - t(W + W|UP-to-W)	**34.98***†	**49.84***†	**34.63***†	**49.39***†
Multi-Source Transformer - t(W + W|UP-to-W|UP)	**34.39***†	**49.62***†	**34.94***†	**50.17***†
Shared-Multi-Source Transformer - t(W + W|UP-to-W)	**34.78***†	**50.05***†	**35.28***†	**50.51***†
Shared-Multi-Source Transformer - t(W + W|UP-to-W|UP)	**35.63***†	**49.96***†	**34.55***†	**50.40***†

Table 8. Experiment results for English-to-Thai translation pair. Bold numbers indicate higher BLEU and chrF scores than the two baseline models. Scores with an asterisk (*) are significantly higher than the first baseline Transformer model and scores with a dagger (†) are significantly higher than the second baseline EDITOR model, at $p < 0.05$.

Models	English-to-Thai by Existing Thai POS Tagger (44 POS tags)		English-to-Thai by New Thai POS Tagger (16 POS tags)	
	BLEU scores (%)	chrF scores (%)	BLEU scores (%)	chrF scores (%)
Baseline Transformer - t(W-W)	31.04	48.81	31.04	48.81
Baseline EDITOR *with soft constraints*	24.60	51.28	24.60	51.28
Transformer- t(W|P-to-W)	31.04†	48.16	31.04†	48.16
Transformer - t(W-to-W|P)	30.92†	48.49	30.16†	47.89
Transformer - t(W|P-to-W|P)	30.68†	47.01	26.42	44.95
Multi-Source Transformer - t(W + W|P-to-W)	**35.50***†	**53.67***†	**35.50***†	**53.67***†
Multi-Source Transformer - t(W + W|P-to-W|P)	**35.67***†	**53.66***†	**34.76***†	**53.15***
Shared-Multi-Source Transformer - t(W + W|P-to-W)	**36.73***†	**54.40***†	**36.73***†	**54.40***†
Shared-Multi-Source Transformer - t(W + W|P-to-W|P)	**35.19***†	**53.16***	**34.98***†	**53.31***
Transformer- t(W|UP-to-W)	30.97†	47.72	30.97†	47.72
Transformer - t(W-to-W|UP)	30.10†	48.18	30.75†	47.95
Transformer - t(W|UP-to-W|UP)	30.77†	47.63	30.67†	47.66
Multi-Source Transformer - t(W + W|UP-to-W)	**36.40***†	**53.92***†	**36.40***†	**53.92***†
Multi-Source Transformer - t(W + W|UP-to-W|UP)	**34.67***†	**53.62***†	**34.41***†	**53.13***
Shared-Multi-Source Transformer - t(W + W|UP-to-W)	**37.03***†	**54.06***†	**37.03***†	**54.06***†
Shared-Multi-Source Transformer - t(W + W|UP-to-W|UP)	**35.50***†	**53.45***†	**34.85***†	**53.47***†

According to the experiment results of our proposed models shown in Tables 4–8, most of the proposed models that used linguistic information, *shared-multi-source transformer models* in particular, significantly outperform the two baseline models. Furthermore, the *transformer* models that integrated linguistic features improved the translation results over the baseline *transformer* model using only word vectors.

When comparing the existing Thai POS tagger to our newly built Thai POS tagger, most of the proposed models that incorporated linguistic features through our newly Thai POS tagger outperformed such models using the existing Thai POS tagger in the Thai-to-Myanmar and Myanmar-to-Thai translations. This is owing to the comparable number of POS tag sets utilized in the Myanmar POS tagger and our newly built Thai POS tagger. Whereas our newly built Thai POS tagger and the Myanmar POS tagger use only 16 and 15 POS tags, respectively, the existing Thai POS tagger uses 44 POS tags. Thus, for the Thai-to-Myanmar and Myanmar-to-Thai translation pairs based on the existing Thai POS tagger, the models have difficulty learning the mapping of the different POS tags during the training process in comparison to the models developed through our newly built Thai POS tagger. Hence, the translation results may be lower than that of the models using the new Thai POS tagger. Although most models utilizing the existing Thai POS tagger and the models that use our newly built Thai POS tagger provide comparable translation performance, we found that *shared-multi-source transformer* models with UPOS tags by our newly built Thai POS tagger {t(W + W|UP-to-W|UP)} outperform such models developed through the existing Thai POS tagger in terms of BLEU scores (+ 0.13) and chrF scores (+ 0.47) for Thai-to-Myanmar, and BLEU scores (+ 0.75) and chrF scores (+ 0.72) for Myanmar-to-Thai translation pairs.

In Thai-to-English translation, except for the *transformer* models integrating UPOS tags, i.e., {t(W|UP-to-W), t(W-to-W|UP), and t(W|UP-to-W|UP)}, all of the remaining proposed models developed through our newly built Thai POS tagger improved the translation performance over such models created using the existing Thai POS tagger. However, most of the proposed models formulated through the existing Thai POS tagger outperformed such models utilizing the new Thai POS tagger for English-to-Thai translation. The numbers of POS tag sets contained in the English POS tagger and the existing Thai POS tagger are comparable, with 45 tags and 44 tags, respectively. The comparable number of POS tag sets makes it easy for the translation models to learn the mapping of different POS tags during training. Thus, the models developed using the existing Thai POS tagger increase the translation performance in comparison to the models using the new Thai POS tagger in English-to-Thai translation.

Unlike the number of POS tag sets contained in the English POS tagger, our newly built Thai POS tagger only includes very few numbers of 16 POS tags. Although the new Thai POS tagger only contains 16 tags, the model developed through this new tagger may learn how to map the different POS tags well when the source side is the Thai language. If the source language is Thai and the translation model uses our newly built Thai POS tagger, the model may easily learn to map the POS tags from the source language to the corresponding POS tags of the target English language during training. As a result, in Thai-to-English translation, most of the proposed models developed through our newly built Thai POS tagger perform better than such models created by the existing Thai POS tagger. In general, the existing Thai POS tagger is more effective than the new Thai POS tagger, particularly for English-to-Thai translation. When the source side is the Thai language or the POS tagger for the target language contains a comparable number of POS tags as in the newly built Thai POS tagger, our newly built Thai POS tagger is more beneficial to the proposed models for Thai-to-Myanmar, Myanmar-to-Thai, and Thai-to-English translation pairs than the existing Thai POS tagger.

The experiment results answer our research questions regarding whether adding linguistic features such as POS or UPOS tags to transformer-based models improves the translation quality for two-way translation of low-resource translation pairs (i.e., Thai–Myanmar, Myanmar–English, and Thai–English). Furthermore, the different configurations of the proposed models provide a translation quality better than or comparable to the baseline *transformer* and EDITOR models.

5.2. Comprehensive analysis

After comparing the experiment results of all our proposed models to the baseline *transformer* and EDITOR models, we present a comprehensive analysis of the translation pairs utilized in our experiments.

In Thai-to-Myanmar translation, all our proposed multi-source models achieve significantly higher BLEU and chrF scores over the first baseline *transformer* model. In addition, the proposed models provide higher significance BLEU scores and better chrF scores than the second baseline EDITOR model. In Thai-to-Myanmar and English-to-Myanmar translations, we notice that the second baseline model yields comparable chrF scores but is lower than our proposed models. During the decoding process, the EDITOR model employs one to three soft lexical constraints, which may make it easier for the model to generate the correct target words. Furthermore, the Myanmar text is made up of Myanmar words with one or more syllables. Each syllable has a number of characters, and chrF scores are also mainly dependent on the n-gram characters. Thus, these may cause the EDITOR model to obtain comparable chrF scores in Thai-to-Myanmar and English-to-Myanmar translations.

Our proposed multi-source models provide better translation performance in English-to-Myanmar translation than those models in the Myanmar-to-English translation. In addition, the EDITOR model improves the translation performance in English-to-Myanmar translation. Myanmar and English languages are extremely distinct languages. Myanmar language has a complicated syntactic structure and constructs the sentences using the Subject-Object-Verb (SOV) order, whereas the English sentences use the Subject-Verb-Object (SVO) order. Thus, the models may have trouble learning the translation from the Myanmar language, which has a complex syntactic structure (i.e., SOV order), to the English language, which follows the SVO order. If the input sentence contains ambiguous words (i.e., one Myanmar word has several English meanings), it may also be difficult for the model to produce the correct target English words in Myanmar-to-English translation. These facts decrease the translation performance of the models in Myanmar-to-English translation compared to those in English-to-Myanmar translation. However, our proposed models that utilize the linguistic features (i.e., POS or UPOS tags) alleviate the problems of ambiguous Myanmar words on the source side and provide significantly higher BLEU and chrF scores over the baseline *transformer* and EDITOR models.

In the result tables, we found that the translation performance of all models in Thai-to-English and English-to-Thai translations yield much higher BLEU and chrF scores than the Thai-to-Myanmar and Myanmar-to-Thai translations. Although Thai and English languages have similar syntax structures (i.e., SVO order), the grammatical order between Thai and Myanmar language pairs is distinct. Thus, the reordering procedure for those language pairs is more complicated, and the translation model underperforms. However, our proposed models outperform the two baseline models and achieve higher BLEU and chrF scores.

In the following sections, we will present different analyzes to prove how our proposed multi-source models are more effective than the baseline *transformer* and EDITOR models. The comparisons between the two baseline models and our proposed multi-source models for Thai-to-Myanmar translation are analyzed using the *compare-mt* [43]. According to our newly built Thai POS tagger, the shared-multi-source transformer model with output (W|POS), namely, *SMulT-WPoS*, and the shared-multi-source transformer model with output (W|UPOS), namely, *SMulT-WUPoS*, were used to compare with the baseline *Transformer* and EDITOR models. The *compare-mt* is a tool used for comparing and analyzing the experiment results of systems applied to language generating tasks, including machine translation. This gives the user a high-level coherent view of the key differences between the systems, which may then be used to lead to further studies or system improvements. The comprehensive analysis of the difference between the two baseline models and shared-multi-source transformer models (*SMulT-WPoS*, and *SMulT-WUPoS*) is mainly described in terms of *bucketed analysis* and *aggregate score analysis*.

5.2.1. Bucketed analysis

The *bucketed analysis* divides words or sentences into buckets and computes salient statistics over these buckets. There are two types of *bucketed analysis*, namely, *word accuracy analysis* and *sentence-level analysis*. The f-measure of the outputs of a system concerning a reference is referred to as the accuracy. The *word accuracy analysis* is used to determine which types of words are generated more on one system than on other system based on the frequency within the training set. The *SMulT-WPoS* and *SMulT-WUPoS* models are more resilient over the first baseline *Transformer* model to low-frequency words on the interval [5,10), whereas *Transformer*, *SMulT-WPoS*, and *SMulT-WUPoS* models provide comparable results on extremely high-frequency words, but lower than the second baseline model, as seen in Fig. 4. Because the second baseline EDITOR model uses lexical constraints, which may benefit producing correct outputs, the EDITOR model performs better than other models on the lowest frequency words. Although the two baseline *Transformer* and EDITOR models, and our proposed *SMulT-WPoS* and *SMulT-WUPoS* models depend on the word frequency in the training dataset to produce the correct words, the *Transformer* and *SMulT-WUPoS* models are unable to perform well on the lowest word frequency count. By contrast, the shared-multi-source transformer model, namely, *SMulT-WPoS*, performs better than the first baseline *Transformer* and *SMulT-WUPoS* models on the lowest word frequency count. Three models (i.e., *Transformer*, *SMulT-WPoS*, and *SMulT-WUPoS*), all yield a comparable word accuracy (i.e., f-measure) on the highest frequency count and are lower than the second baseline EDITOR model. According to this *word accuracy analysis*, the first baseline *Transformer* model is more data-hungry than the second baseline EDITOR model and the shared-multi-source transformer models (i.e., *SMulT-WPoS* and *SMulT-WUPoS*) in generating correct words.

The *sentence-level analysis* is applied to determine which types of sentences on one system work better than on the other. The evaluation was conducted using three different bucket and statistic types: the statistic is the *score* based on the *sentence length* bucket, the static is the *count* based on the difference between the output and reference lengths {*len(output)* − *len(reference)*} bucket, and the statistic is the *count* based on the *sentence-level*

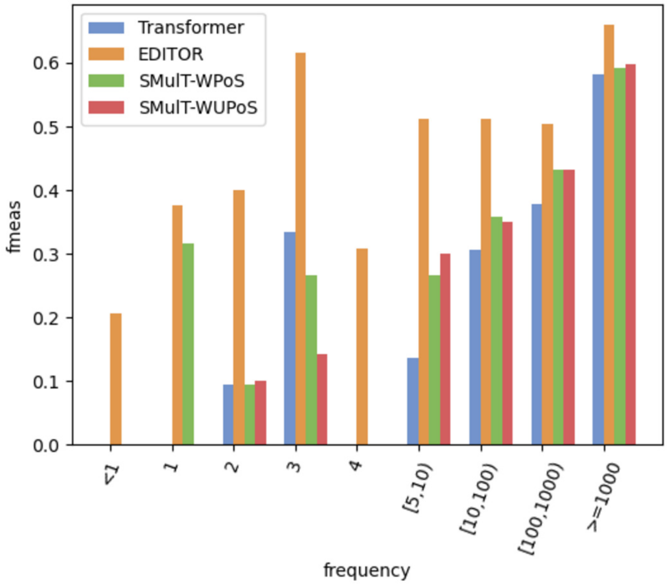


Fig. 4. Word accuracy analysis by frequency bucket in the training set.

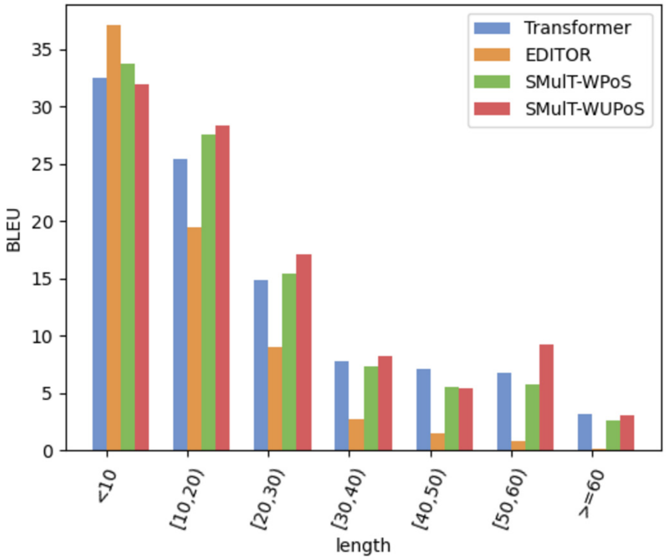


Fig. 5. BLEU scores based on sentence length bucket.

BLEU score bucket. In the case of (bucket = length, statistic = score), this determines the BLEU score based on the reference sentence length, showing whether a system performs better or worse when sentences are shorter or longer. The sentence lengths contained in the dataset can affect the translation performance of the models. Fig. 5 shows the measured BLEU scores based on the sentence buckets of the baseline *Transformer* and EDITOR models, and shared-multi-source transformer models (*SMulT-WPoS, and SMulT-WUPoS*). As shown in Fig. 5, the second baseline EDITOR model and *SMulT-WPoS* model achieve better results for extremely short sentences, whereas the *Transformer* and *SMulT-WUPoS* models perform better for extremely long sentences. Based on this point, we can see that the EDITOR and the *SMulT-WPoS* models, unlike the *Transformer* and *SMulT-WUPoS* models, are unable to perform well on long sentences. One of our proposed models with UPOS tags, i.e., the *SMulT-WUPoS* model, provides higher BLEU scores than the two baseline models and *SMulT-WPoS* model when the sentence length is between 10 to 20 words. Furthermore, the *SMulT-WUPoS* model significantly outperforms the second baseline EDITOR model and may compete with the first baseline *Transformer* model to provide a better translation performance on long sentences.

If the bucket and statistic types are *lengthdiff* and *count*, respectively, this produces the number of sentences that have a specific difference in length between the reference and the output. The closer the length difference is to zero, the more likely is the system able to match the output length; however, the flatter the difference is, the more difficult it is for the system to generate the correct sentence length. We can observe in Fig. 6 that the *SMulT-WUPoS* model generates a more accurate sentence length than the *Transformer*, EDITOR and *SMulT-WPoS* models. The correct sentence length has a direct impact on the increase in translation quality. Because our proposed shared-multi-source transformer model with UPOS tags, the *SMulT-WUPoS* model, generates the most accurate sentence count, this model may yield the best translation quality in comparison with the baseline *Transformer* and EDITOR models. Thus, our proposed *SMulT-WUPoS* model outperforms the two baseline models for Thai-to-Myanmar translation.

In the case of (bucket = score, statistic = count), this produces a number of sentences obtaining a specific score (i.e., sentence-level BLEU score). This indicates how many sentences each system can generate with a certain accuracy. The first baseline *Transformer* model produces more

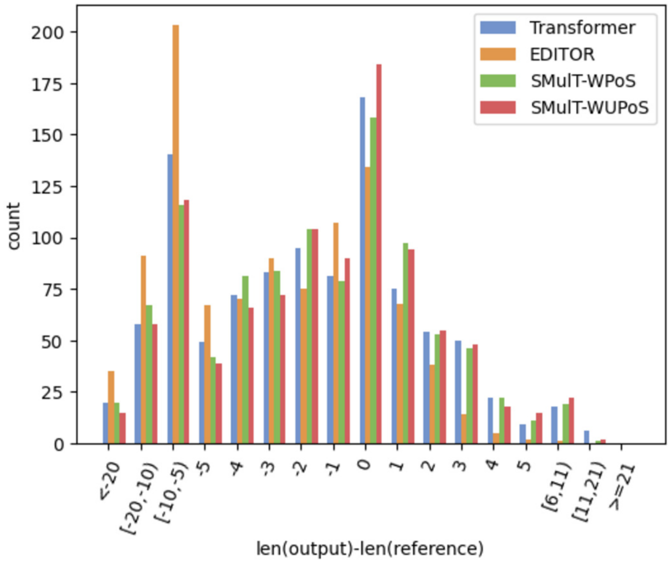


Fig. 6. Counts of sentences based on difference in length between the reference and the output.

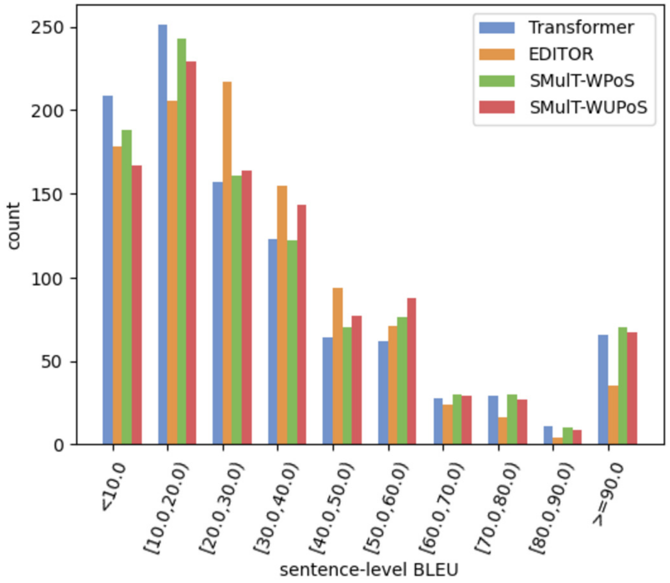


Fig. 7. Counts of sentences based on sentence-level BLEU bucket.

sentences with low sentence-level BLEU scores (i.e., < 10.0) than the remaining three models, whereas the *SMulT-WPoS* model generates more sentences with higher scores (i.e., >= 90.0) than the two baseline *Transformer* and EDITOR models, and the *SMulT-WUPoS* model, as seen in Fig. 7. The number of optimal sentence-level BLEU scores directly affects the translation quality produced by the model. Thus, the higher the number of optimal sentence-level BLEU scores, the better the translation performance. Between the sentence-level BLEU scores of 40.0 and 50.0, the second baseline model yields a higher sentence count. However, according to Fig. 7, our proposed *SMulT-WUPoS* model produces a more sentence count between sentence-level BLEU scores of 50.0 and 60.0 than the two baseline models. Furthermore, the *SMulT-WPoS* model generates more sentences with sentence-level BLEU scores of 90.0 and beyond 90.0 than the baseline *Transformer* and EDITOR models. Owing to the higher sentence count of the optimal sentence-level BLEU scores provided by the *SMulT-WPoS* model, our proposed *SMulT-WPoS* model increases the translation quality over the two baseline models for Thai-to-Myanmar translation.

5.2.2. *Aggregate score analysis*

The *aggregate score analysis* utilizes the standard BLEU [47] as a default evaluation metric, computed using the entire test data, and determines which system is more accurate overall. BLEU is an evaluation metric used to measure the difference between reference translations (i.e., translations created by humans) and automatic translations (i.e., hypothesis produced by the translation models) of the same source data. The results of the *aggregate score analysis* on the baseline *Transformer* and EDITOR models, and our proposed *SMulT-WPoS* and *SMulT-WUPoS* models are shown in Fig. 8. The results demonstrate that the *SMulT-WUPoS* model outperforms the first baseline *Transformer* model (+2.35), the second baseline EDITOR model (+7.11), and the *SMulT-WPoS* model (+0.88) in terms of BLEU scores. The BLEU score determines the quality of a translation, with a higher BLEU score indicates a better translation quality. Thus, our proposed *SMulT-WUPoS* model performs well on Thai-to-Myanmar translation over

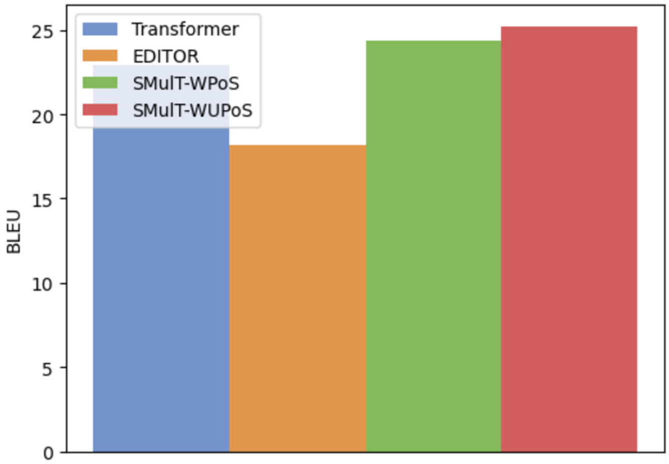


Fig. 8. Aggregate score analysis based on standard BLEU score.

Table 9. The test set unigram F1 scores of occurrence in the predicted sentences based on the frequencies in the training corpus for multi-source transformer and shared-multi-source transformer models with POS and UPOS tags for Thai-to-Myanmar translation.

Word Freq	MulT-WPoS	MulT-WUPoS	SMulT-WPoS	SMulT-WUPoS
1	**0.35**	0.33	**0.32**	0.00
2	**0.19**	0.09	0.10	0.10
3	0.13	**0.17**	**0.27**	0.14
4	0.00	0.00	0.00	0.00
[5, 10)	0.27	**0.29**	0.27	**0.30**
[10, 100)	0.36	0.36	**0.36**	0.35
[100, 1000)	0.43	**0.44**	0.43	0.43
≧1000	0.59	**0.60**	0.59	**0.60**

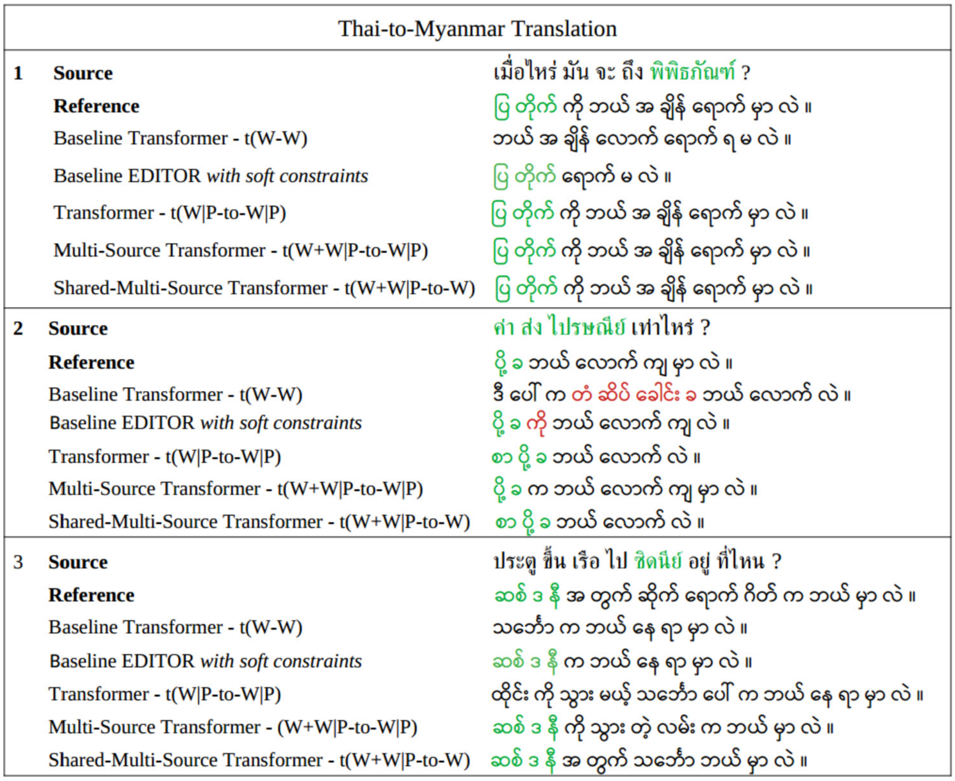
the baseline *Transformer* and EDITOR models. According to both *bucketed analysis* and *aggregate score analysis*, we can conclude that our proposed *SMulT-WPoS* and *SMulT-WUPoS* models outperform the baseline *Transformer* and EDITOR models for Thai-to-Myanmar translation.

5.3. Analysis between the models that applied POS and UPOS tags

In this section, we analyze the differences between the proposed models with POS and UPOS tags for Thai-to-Myanmar translation. For the analysis of differences between the models that utilize POS and UPOS tags, we compared the multi-source transformer model with POS tags (*MulT-WPoS*) to the multi-source transformer model with UPOS tags (*MulT-WUPoS*). Moreover, we also compared the shared-multi-source transformer model with POS tags (*SMulT-WPoS*) to the shared-multi-source transformer model with UPOS tags (*SMulT-WUPoS*). Table 9 shows how f-measure (F1 score) varies with the word frequency. The multi-source transformer and shared-multi-source transformer with POS tags, namely *MulT-WPoS* and *SMulT-WPoS* models, improve translation accuracy more than the multi-source transformer and shared-multi-source transformer with UPOS tags, namely *MulT-WUPoS* and *SMulT-WUPoS* models, particularly for low-frequency and rare words. For the high-frequency words, *MulT-WUPoS* and *SMulT-WUPoS* models perform better than the multi-source transformer and shared-multi-source transformer models with POS tags (i.e., *MulT-WPoS* and *SMulT-WPoS*). The UPOS tags that are applied in the translation model are identical in both the source and target languages for the translation pairs, and contain fewer tags than the original POS tags, enabling the model to generate more accurate words. Thus, most of our proposed models with UPOS tags outperform the models with POS tags in general.

5.4. Translation errors and analysis

Samples of Thai-to-Myanmar and Myanmar-to-Thai translations produced by the proposed models are listed in Figs. 9 and 10. Fig. 9 shows three sample translations of our proposed models, i.e., the *transformer* model with POS tags {t(W|P-to-W|P)}, *multi-source transformer* model with POS tags {t(W + W|P-to-W|P)} and *shared-multi-source transformer* with POS tags {t(W+ W|P-to-W)}, and the baseline *transformer* {t(W-to-W)} and EDITOR models on the task of Thai-to-Myanmar translation. In this figure, we list those models that produce the best translation result with linguistic features according to our newly built Thai POS tagger for Thai-to-Myanmar translation from among all of the proposed models. We hoped to determine how linguistic features aid in solving problems that appear in the baseline translation systems. For the first sample in Fig. 9-(1), our proposed models with POS tags obtain the correct translation "ပြ တိုက်" (museum in English) for the word "พิพิธภัณฑ์" whereas the translation for the word in the first baseline *transformer* model is missing. Although the correct word "ပြ တိုက်" can be produced by the second baseline EDITOR model, it cannot generate the whole sentence correctly compared to our proposed models. We can see from this example that POS tags can help address the lack of meaningful words. In Fig. 9-(2), the *multi-source transformer* model with POS tags {t(W+ W|P-to-W|P)} can manage to translate the source phrase "ค่า ส่ง ไปรษณีย์" ("postage" in English) into "ပို့ ခ" with exactly the correct word order, whereas the baseline *transformer* model translates the incorrect word "တံ ဆိပ် ခေါင်း ခ" ("stamp charge" in English) instead of the word "ပို့ ခ." Moreover, the second baseline EDITOR model can produce the correct translation "ပို့ ခ," but it generates extra incorrect word "ကို." The *transformer* model with POS tags {t(W|P-to-W|P)} and the *shared-multi-source transformer* model with POS tags {t(W+ W|P-to-W)} can also translate a similar meaning of the correct Myanmar word (i.e., the words "ပို့ ခ" and



Thai-to-Myanmar Translation

1	**Source**	เมื่อไหร่ มัน จะ ถึง พิพิธภัณฑ์ ?
	Reference	ပြ တိုက် ကို ဘယ် အ ချိန် ရောက် မှာ လဲ ။
	Baseline Transformer - t(W-W)	ဘယ် အ ချိန် လောက် ရောက် ရ မ လဲ ။
	Baseline EDITOR *with soft constraints*	ပြ တိုက် ရောက် မ လဲ ။
	Transformer - t(W|P-to-W|P)	ပြ တိုက် ကို ဘယ် အ ချိန် ရောက် မှာ လဲ ။
	Multi-Source Transformer - t(W+W|P-to-W|P)	ပြ တိုက် ကို ဘယ် အ ချိန် ရောက် မှာ လဲ ။
	Shared-Multi-Source Transformer - t(W+W|P-to-W)	ပြ တိုက် ကို ဘယ် အ ချိန် ရောက် မှာ လဲ ။
2	**Source**	ค่า ส่ง ไปรษณีย์ เท่าไหร่ ?
	Reference	ပို့ ခ ဘယ် လောက် ကျ မှာ လဲ ။
	Baseline Transformer - t(W-W)	ဒီ ပေါ် က တံ ဆိပ် ခေါင်း ခ ဘယ် လောက် လဲ ။
	Baseline EDITOR *with soft constraints*	ပို့ ခ ကို ဘယ် လောက် ကျ လဲ ။
	Transformer - t(W|P-to-W|P)	စာ ပို့ ခ ဘယ် လောက် လဲ ။
	Multi-Source Transformer - t(W+W|P-to-W|P)	ပို့ ခ က ဘယ် လောက် ကျ မှာ လဲ ။
	Shared-Multi-Source Transformer - t(W+W|P-to-W)	စာ ပို့ ခ ဘယ် လောက် လဲ ။
3	**Source**	ประตู ขึ้น เรือ ไป ซิดนีย์ อยู่ ที่ไหน ?
	Reference	ဆစ် ဒ နီ အ တွက် ဆိုက် ရောက် ဂိတ် က ဘယ် မှာ လဲ ။
	Baseline Transformer - t(W-W)	သင်္ဘော က ဘယ် နေ ရာ မှာ လဲ ။
	Baseline EDITOR *with soft constraints*	ဆစ် ဒ နီ က ဘယ် နေ ရာ မှာ လဲ ။
	Transformer - t(W|P-to-W|P)	ထိုင်း ကို သွား မယ့် သင်္ဘော ပေါ် က ဘယ် နေ ရာ မှာ လဲ ။
	Multi-Source Transformer - (W+W|P-to-W|P)	ဆစ် ဒ နီ ကို သွား တဲ့ လမ်း က ဘယ် မှာ လဲ ။
	Shared-Multi-Source Transformer - t(W+W|P-to-W)	ဆစ် ဒ နီ အ တွက် သင်္ဘော ဘယ် မှာ လဲ ။

Fig. 9. Sample translations of the proposed models and the two baseline models on Thai-to-Myanmar translation. Words in green indicate a correct translation. Red words are incorrect translations for the baseline models.

"စာ ပို့ ခ" generally have the same meaning in Myanmar speaking form). In Fig. 9-(3), the source word "ซิดนีย์" is a city name meaning "Sydney" in English, and the baseline *transformer* model misses translating the source word "ซิดนีย์," whereas the multi-source models with POS tags and the second baseline EDITOR model achieve the correct translation "ဆစ် ဒ နီ" for it. Although the second baseline model can translate the correct word, it cannot manage to provide the correct postposition "အ တွက်" ("for" in English) followed by the word "ဆစ် ဒ နီ." The integration of linguistic features into NMT models helps in translating the correct city name and postposition. These translation samples demonstrate that our proposed models with POS tags are more capable of generating a correct sentence structure than the two baseline models.

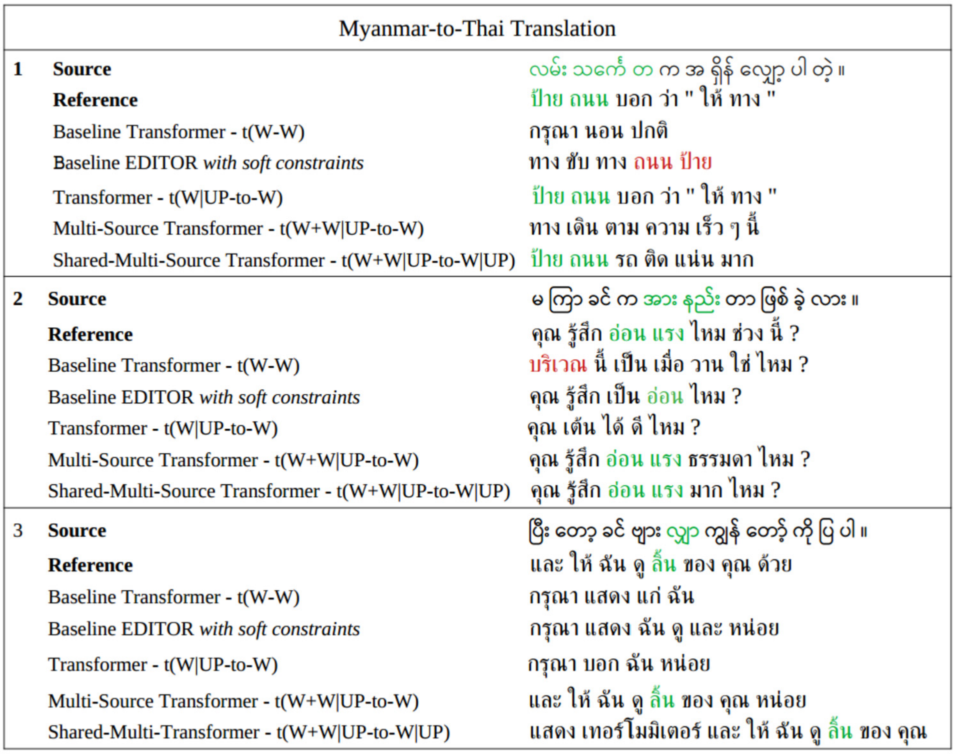
Fig. 10 shows three sample Myanmar-to-Thai translations. In this figure, we also selected the models that yield the best translation result for each proposed model, i.e., the *transformer* model with UPOS tags {t(W|UP-to-W)}, the *multi-source transformer* model with UPOS tags {t(W + W|UP-to-W)}, and the *shared-multi-source transformer* model with UPOS tags {t(W+ W|UP-to-W|UP)} for comparison with the baseline *transformer* and EDITOR models on Myanmar-to-Thai translation. In Fig. 10-(1), the source word "လမ်း သင်္ကေတ" ("road sign" in English) receives a perfect translation "ป้าย ถนน" ("road sign" in English) when using the *transformer* model with UPOS tags {t(W|UP-to-W)} and the *shared-multi-source transformer* with UPOS tags {t(W+ W|UP-to-W|UP)}, whereas its translation when applying the first baseline *transformer* model is missing. Furthermore, the second baseline EDITOR model gives the wrong word order "ถนน ป้าย" for the correct phrase "ป้าย ถนน" ("road sign" in English). In this translation, the two baseline models are unable to produce the whole sentence correctly when compared to our proposed *transformer* model with UPOS tags {t(W|UP-to-W)}. In Fig. 10-(2), the source word "အား နည်း" means "weak" in English, and our proposed *multi-source transformer* and *shared-multi-source transformer* models with UPOS tags achieve the correct translation "อ่อน แรง," whereas the first baseline *transformer* model translates the incorrect word "บริเวณ" ("area" in English). In addition, the second baseline EDITOR model translates only the correct subword "อ่อน" for the complete translation "อ่อน แรง," and is missing to translate the remaining subword "แรง." In Fig. 10-(3), our *multi-source transformer* and *shared-multi-source transformer* models with UPOS tags obtain the correct translation "ลิ้น" ("tongue" in English), whereas the baseline *transformer* and EDITOR models miss translating the source word "လျှာ." Figs. 9 and 10 show that our proposed models with linguistic features can produce better translation results with a more correct word selection and sentence structure than the first baseline *transformer* model utilizing only word vectors and the second baseline EDITOR model. Thus, our proposed models with POS or UPOS tags outperformed the two baseline models on Thai-to-Myanmar and Myanmar-to-Thai translations.

5.5. Word errors analysis

The Word Error Rate (*WER*) was used to analyze the translated outputs of our proposed models and the two baseline models (*transformer* and EDITOR). Thai-to-English and English-to-Thai translation pairs were selected to compare the *WER* between the models. We used the SCLITE program version 2.10 of the SCTK toolkit (https://github.com/usnistgov/SCTK) [49] for calculating the *WER* based on the alignments between the reference and hypothesis. The formula for *WER* can be stated as the equation (3):

$$WER = \frac{(I + D + S) \times 100}{N} \tag{3}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correct words and N is the number of words in the reference ($N = S + D + C$). When the number of insertions is very high, the percentage of WER can be greater than 100%.



Myanmar-to-Thai Translation		
1	**Source**	လမ်း သင်္ကေ တ က အ ရှိန် လျှော့ ပါ တဲ့ ။
	Reference	ป้าย ถนน บอก ว่า " ให้ ทาง "
	Baseline Transformer - t(W-W)	กรุณา นอน ปกติ
	Baseline EDITOR *with soft constraints*	ทาง ชับ ทาง ถนน ป้าย
	Transformer - t(W|UP-to-W)	ป้าย ถนน บอก ว่า " ให้ ทาง "
	Multi-Source Transformer - t(W+W|UP-to-W)	ทาง เดิน ตาม ความ เร็ว ๆ นี้
	Shared-Multi-Source Transformer - t(W+W|UP-to-W|UP)	ป้าย ถนน รถ ติด แน่น มาก
2	**Source**	မ ကြာ ခင် က အား နည်း တာ ဖြစ် ခဲ့ လား ။
	Reference	คุณ รู้สึก อ่อน แรง ไหม ช่วง นี้ ?
	Baseline Transformer - t(W-W)	บริเวณ นี้ เป็น เมื่อ วาน ใช่ ไหม ?
	Baseline EDITOR *with soft constraints*	คุณ รู้สึก เป็น อ่อน ไหม ?
	Transformer - t(W|UP-to-W)	คุณ เต้น ได้ ดี ไหม ?
	Multi-Source Transformer - t(W+W|UP-to-W)	คุณ รู้สึก อ่อน แรง ธรรมดา ไหม ?
	Shared-Multi-Source Transformer - t(W+W|UP-to-W|UP)	คุณ รู้สึก อ่อน แรง มาก ไหม ?
3	**Source**	ပြီး တော့ ခင် ဗျား လျှာ ကျွန် တော့် ကို ပြ ပါ ။
	Reference	และ ให้ ฉัน ดู ลิ้น ของ คุณ ด้วย
	Baseline Transformer - t(W-W)	กรุณา แสดง แก่ ฉัน
	Baseline EDITOR *with soft constraints*	กรุณา แสดง ฉัน ดู และ หน่อย
	Transformer - t(W|UP-to-W)	กรุณา บอก ฉัน หน่อย
	Multi-Source Transformer - t(W+W|UP-to-W)	และ ให้ ฉัน ดู ลิ้น ของ คุณ หน่อย
	Shared-Multi-Transformer - t(W+W|UP-to-W|UP)	แสดง เทอร์โมมิเตอร์ และ ให้ ฉัน ดู ลิ้น ของ คุณ

Fig. 10. Sample translations of the proposed models and the two baseline models on Myanmar-to-Thai translation. Words in green indicate a correct translation. Red words are incorrect translations for the baseline models.

Table 10. Comparison of *WER* scores (lower is the better) for Thai-to-English and English-to-Thai translation pairs. Lower *WER* scores than the two baseline *transformer* and EDITOR models are highlighted as bold numbers. The scores marked in bold with an underline are the lowest *WER* scores when compared to the two baseline models.

Models	Thai-to-English		English-to-Thai	
	WER (%) by Existing Thai POS Tagger	WER (%) by New Thai POS Tagger	WER (%) by Existing Thai POS Tagger	WER (%) by New Thai POS Tagger
Baseline Transformer - t(W-W)	54.80	54.80	62.40	62.40
Baseline EDITOR *with soft constraints*	54.40	54.40	60.60	60.60
Transformer- t(W|P-to-W)	54.90	55.30	62.50	62.50
Transformer - t(W-to-W|P)	55.20	55.20	62.30	63.70
Transformer - t(W|P-to-W|P)	55.10	54.90	63.40	64.70
Multi-Source Transformer - t(W + W|P-to-W)	**49.50**	**49.30**	**55.90**	**55.90**
Multi-Source Transformer - t(W + W|P-to-W|P)	**50.70**	**51.20**	**55.30**	**56.30**
Shared-Multi-Source Transformer - t(W + W|P-to-W)	**50.60**	**49.30**	**53.80**	**53.80**
Shared-Multi-Source Transformer - t(W + W|P-to-W|P)	**50.80**	**49.50**	**56.40**	**55.90**
Transformer- t(W|UP-to-W)	55.20	55.60	63.70	63.70
Transformer - t(W-to-W|UP)	54.70	54.70	62.00	63.10
Transformer - t(W|UP-to-W|UP)	53.70	57.50	61.80	62.20
Multi-Source Transformer - t(W + W|UP-to-W)	**49.00**	**50.60**	**55.10**	**55.10**
Multi-Source Transformer - t(W + W|UP-to-W|UP)	**50.50**	**50.60**	**54.90**	**55.40**
Shared-Multi-Source Transformer - t(W + W|UP-to-W)	**50.00**	**50.00**	**54.10**	**54.10**
Shared-Multi-Source Transformer - t(W + W|UP-to-W|UP)	**49.10**	**50.70**	**55.60**	**56.00**

Table 10 shows the comparison of *WER* scores between our proposed models and the two baseline *transformer* and EDITOR models. In this table, all our proposed multi-source models achieved the lower *WER* scores than the two baseline models for both Thai-to-English and English-to-Thai translation pairs. When we compared the lowest scores of the proposed multi-source models to the baseline models, the *multi-source transformer* model with UPOS tags using the existing Thai POS tagger {t(W + W|UP-to-W)} provides the lowest *WER* scores (−5.80) over the first baseline *transformer* model and (−5.40) over the second baseline EDITOR model in Thai-to-English translation. Moreover, in English-to-Thai translation, our proposed *shared-multi-source transformer* models with POS tags {t(W+ W|P-to-W)} achieve the lowest *WER* scores (−8.60) and (−6.80) over the first baseline *transformer* and second baseline EDITOR models. The lower the *WER* scores, the better the translation quality of the models. Thus, our proposed multi-source models perform better than the two baseline models.

We notice that the multi-source models with output POS tags (W|P), i.e., {t(W + W|P-to-W|P)} also provide lower WER scores than the two baseline models that use only word vectors on the source and target sides (i.e., {t(W-W)}), and such models outperform the baseline models. We made a manual error analysis of the models in Thai-to-English translation to know how effective the POS tags are on the target side of the model. For this, our proposed *shared multi-source transformer* model with output POS tags (W|P) using our newly built Thai POS tagger, and the two baseline models, were used for conducting the analysis. We describe the analysis samples in Tables 11–13.

Table 11. Sample 1 for detailed calculations of word error analysis. The capital words highlight the number of insertions, deletions, and substitutions.

Baseline Transformer - t(W-W):
Scores: (#C #S #D #I) 7 0 2 0
REF: where do you want to check YOUR LUGGAGE?
HYP: where do you want to check **** *******?
Eval: D D
Baseline EDITOR with *soft constraints:*
Scores: (#C #S #D #I) 6 1 2 0
REF: where do you want TO CHECK YOUR luggage?
HYP: where do you want ** ***** THE luggage?
Eval: D D S
Shared-Multi-Source Transformer - t(W + W|P-to-W|P):
Scores: (#C #S #D #I) 7 2 0 0
REF: where do you want to CHECK YOUR luggage?
HYP: where do you want to EXAMINE THE luggage?
Eval: S S

Table 12. Sample 2 for detailed calculations of word error analysis. The capital words highlight the number of insertions, deletions, and substitutions.

Baseline Transformer - t(W-W):
Scores: (#C #S #D #I) 4 2 0 1
REF: first * CLASS or SECOND class?
HYP: first, SECOND or THIRD class?
Eval: I S S
Baseline EDITOR with *soft constraints:*
Scores: (#C #S #D #I) 5 0 1 1
REF: first CLASS or ****** second class?
HYP: first ***** or SECOND second class?
Eval: D I
Shared-Multi-Source Transformer - t(W + W|P-to-W|P):
Scores: (#C #S #D #I) 6 0 0 0
REF: first class or second class?
HYP: first class or second class?
Eval:

Table 13. Sample 3 for detailed calculations of word error analysis. The capital words highlight the number of insertions, deletions, and substitutions.

Baseline Transformer - t(W-W):
Scores: (#C #S #D #I) 12 0 0 1
REF: how long does it take me to get there by ******* train?
HYP: how long does it take me to get there by EXPRESS train?
Eval: I
Baseline EDITOR with *soft constraints:*
Scores: (#C #S #D #I) 10 1 1 0
REF: how long does it take me TO get THERE by train?
HYP: how long does it take me ** get BY by train?
Eval: D S
Shared-Multi-Source Transformer - t(W + W|P-to-W|P):
Scores: (#C #S #D #I) 12 0 0 0
REF: how long does it take me to get there by train?
HYP: how long does it take me to get there by train?
Eval:

In the first sample presented in Table 11, we found that the *object* missing (“YOUR LUGGAGE”) in the hypothesis generated by the first baseline *transformer* model. The *number of insertions* I = 0, *number of deletions* D = 2, *number of substitutions* S = 0, *number of correct words* C = 7, and *number of words in the reference* N = 9 happen, and its WER score is 22.22%. The *preposition* and *verb* missing (“TO CHECK”) occur in the hypothesis produced by the second baseline EDITOR model, which has a WER score of 33.33%. Our proposed *shared-multi-source transformer* model with output POS tags, i.e., {W + W|P-to-W|P} can generate the hypothesis more correctly than the two baseline models without missing any words. In the second sample shown in Table 12, the baseline *transformer* model generates the hypothesis with incorrect translations and it provides the WER score of 50%. Moreover, the second baseline EDITOR model gives a WER score of 33.33% and produces the hypothesis with one word missing (“CLASS”) and one extra word (“SECOND”). However, the proposed *shared-multi-source transformer* model generates the correct hypothesis with a WER score of 0%. In this second sample, only our proposed model can produce the whole sentence correctly compared to the two baseline models. In the last sample described in Table 13, the first baseline *transformer* model output the correct hypothesis, but it produces one extra word (“EXPRESS”) with a WER score of 8.30%. The *preposition* missing is found in the hypothesis produced by the second baseline EDITOR model, which has a WER score of 18.18%. However, our proposed *shared-multi-source transformer* model can generate the whole sentence correctly. According to these samples shown

Table 14. Top 10 confusion pairs of Thai-to-English translation. Ref and Hyp mean the reference and hypothesis, respectively.

Baseline Transformer - t(W-W)		Baseline EDITOR *with soft constraints*		Shared-Multi-Source Transformer - t(W + W|P-to-W|P)	
Frequency	Ref → Hyp	Frequency	Ref → Hyp	Frequency	Ref → Hyp
11	the → a	8	a → the	8	are → is
6	like → want	4	does → is	7	like → want
8	are → is	4	the → a	6	want → like
5	for → to	4	ticket → tickets	5	can → may
5	want → like	3	for → to	5	the → a
4	can → may	3	do → would	4	a → an
3	a → an	3	like → want	4	can → could
3	an → a	3	luggage → suitcase	4	have → do
3	can → could	3	are → is	3	a → the
3	could → can	3	in → to	2	certainly → sure

in Tables 11–13, our proposed *shared-multi-source transformer* model with output POS tags (W|P) performs better than the two baseline models and generates the hypothesis sentences without missing any words. Thus, utilizing the POS tags (W|P) on the target side is more advantageous to the model to generate the correct hypothesis than applying the word vectors (W) on the target side.

In Table 14, we extracted the examples of the top 10 confusion pairs of Thai-to-English translation from one of our proposed multi-source models (i.e., *shared-multi-source transformer* with POS tags by the new Thai POS tagger {t(W + W|P-to-W|P)}) and the two baseline *transformer* and EDITOR models. In this table, we found that most confusion pairs such as "*are* → *is* and *ticket* → *tickets*," and "*a* → *the* and *the* → *a*" are caused by the lack of plural forms and articles in the Thai language. The word pairs such as "*like* and *want*," "*luggage* and *suitcase*," "*for* and *to*," and "*certainly* and *sure*" have similar meanings in the Thai language. The aforementioned words cause confusion in Thai-to-English translation and decrease the translation quality. If the confusion pairs can be cleaned out, the translation performance will improve.

6. Conclusion

In this study, we proposed NMT systems that incorporate linguistic features for the translation of Thai-to-Myanmar, Myanmar-to-English, Thai-to-English, and vice versa. The transformer-based models, such as the transformer models, multi-source transformer models, and shared-multi-source transformer models, were utilized to experiment on each language pair. The POS tagged data were needed to apply NMT models incorporating linguistic features. For this, we used RDRPOSTagger for Thai and English data and the Myanmar POS tagger that we developed in this study for Myanmar data. We also built a new Thai POS tagger in addition to the RDRPOSTagger for Thai data, and both Thai POS taggers were used and experimented with for each model. Several configurations were conducted to evaluate the performance of NMT models integrating linguistic features. The transformer model that utilizes the word vectors and the EDITOR model with soft constraints were used as the baseline models. Our results show that integrating linguistic features in NMT models helps to improve the translation performance for all language pairs. The proposed models achieved better translation quality compared with the baseline *transformer* and EDITOR models. The best model (i.e., shared-multi-source transformer model) yields the highest significant BLEU and chrF scores for two-way translation of all low-resource language pairs.

To the best of our knowledge, this study can be regarded as the first work on the factored neural machine translation that utilizes POS tag information for the low-resource language pairs, namely, Thai–English and Thai–Myanmar. In the future, we intend to incorporate the dependency information into transformer and sequence-to-sequence architecture-based models for low-resource languages.

Declarations

Author contribution statement

Zar Zar Hlaing: Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Ye Kyew Thu: Conceived and designed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Thepchai Supnithi: Conceived and designed the experiments.

Ponrudee Netisopakul: Conceived and designed the experiments; Contributed reagents, materials, analysis tools or data; Analyzed and interpreted the data; Wrote the paper.

Funding statement

Zar Zar Hlaing was supported by King Mongkut's Institute of Technology Ladkrabang [KDS2019/030].

Data availability statement

Data associated with this study has been deposited at https://github.com/ye-kyaw-thu/myPOS/tree/master/myPOS-to-universalPOS.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Acknowledgements

We would like to thank anonymous reviewers and the section editor for their thorough review and for providing valuable comments to improve our paper. We greatly appreciate National Electronics and Computer Technology Center (NECTEC) Thailand for sharing ASEAN-MT Corpus.

Appendix A

See Tables A.1, A.2, A.3, A.4 and A.5.

Table A.1. Universal part-of-speech (UPOS) tags [40] and description.

No.	UPOS Tag	Description
1	VERB	Verbs (all tenses and modes)
2	NOUN	Nouns (common and proper)
3	PRON	Pronouns
4	ADJ	Adjectives
5	ADV	Adverbs
6	ADP	Adpositions (prepositions and postpositions)
7	CONJ	Conjunctions
8	DET	Determiners
9	NUM	Cardinal numbers
10	PRT	Particles or other function words
11	X	Other: foreign words, typos, and abbreviations
12	.	Punctuation

Table A.2. Mapping scheme between POS tags and universal POS tags of Thai language (based on existing Thai POS tagger that used 44 POS tags).

No.	POS Tag [38]	Description	Corresponding UPOS Tag [40]
1	NCMN	Common noun	NOUN
2	NTTL	Title noun	NOUN
3	CNIT	Unit classifier	NOUN
4	CLTV	Collective classifier	NOUN
5	CMTR	Measurement classifier	NOUN
6	CFQC	Frequency classifier	NOUN
7	CVBL	Verbal classifier	NOUN
8	VACT	Active verb	VERB
9	VSTA	Stative verb	VERB
10	NPRP	Proper noun	NOUN
11	NONM	Ordinal number	ADJ
12	VATT	Attributive verb	ADJ
13	DONM	Determiner, ordinal number expression	ADJ
14	ADVN	Adverb with normal form	ADV
15	ADVI	Adverb with iterative form	ADV
16	ADVP	Adverb with prefixed form	ADV
17	ADVS	Sentential adverb	ADV
18	PPRS	Personal pronoun	PRON
19	PDMN	Demonstrative pronoun	PRON
20	PNTR	Interrogative pronoun	PRON
21	DDAN	Definite determiner, after noun without classifier in between	DET
22	DDAC	Definite determiner, allowing classifier in between	DET
23	DDAQ	Definite determiner, following quantitative expression	DET
24	DDBQ	Definite determiner, between noun and classifier or preceding quantitative expression	DET
25	DIAC	Indefinite determiner, following noun; allowing classifier in between	DET
26	DIBQ	Indefinite determiner, between noun and classifier or preceding quantitative expression	DET
27	DIAQ	Indefinite determiner, following quantitative expression	DET
28	NCNM	Cardinal number	NUM
29	NLBL	Label noun	NUM
30	DCNM	Determiner, cardinal number expression	NUM
31	XVBM	Pre-verb auxiliary, before negator	VERB
32	XVAM	Pre-verb auxiliary, after negator	VERB
33	XVMM	Pre-verb, before or after negator	VERB
34	XVAE	Post-verb auxiliary	VERB
35	RPRE	Preposition	ADP
36	JCRG	Coordinating conjunction	CONJ

(*continued on next page*)

Table A.2 (*continued*)

No.	POS Tag [38]	Description	Corresponding UPOS Tag [40]
37	PREL	Relative pronoun	CONJ
38	JSBR	Subordinating conjunction	CONJ
39	JCMP	Comparative conjunction	CONJ
40	FIXN	Nominal prefix	PRT
41	FIXV	Adverbial prefix	PRT
42	EITT	Ending for interrogative sentence	PRT
43	NEG	Negator	PRT
44	PUNC	Punctuation	.

Table A.3. Mapping scheme between POS tags and universal POS tags of Thai language (based on new Thai POS tagger that used 16 POS tags).

No.	POS Tag [37]	Description	Corresponding UPOS Tag [40]
1	AJ	Attribute, modifier, or description of a noun	ADJ
2	AV	Word that modifies or qualifies an adjective, verb, or another adverb	ADV
3	AX	Tense, aspect, mood, and voice	VERB
4	CC	Conjunction and relative pronoun	CONJ
5	CL	Class or measurement unit to which a noun or an action belongs	NOUN
6	FX	Inflectional (nominalizer, adjectivizer, adverbializer, and courteous verbalizer), and derivational	NOUN
7	IJ	Exclamation word	X
8	NG	Word of negation	PRT
9	NN	Person, place, thing, abstract concept, and proper name	NOUN
10	NU	Quantity for counting and calculation	NUM
11	PA	Politeness, intention, belief, question	PRT
12	PR	Word used to refer to an element in the discourse	PRON
13	PS	Location, comparison, instrument, exemplification	ADP
14	PU	Punctuation mark	.
15	VV	Action, state, occurrence, and word that forms the predicate part	VERB
16	XX	Unknown category	X

Table A.4. Mapping scheme between POS tags and Universal POS tags of Myanmar language.

No.	POS Tag [36]	Description	Corresponding UPOS Tag [40]
1	abb	Abbreviation	X
2	adj	Adjective	ADJ
3	adv	Adverb	ADV
4	conj	Conjunction	CONJ
5	fw	Foreign word	X
6	int	Interjection	X
7	n	Noun	NOUN
8	num	Number	NUM
9	part	Particle	PRT
10	ppm	Post positional marker	ADP
11	pron	Pronoun	PRON
12	punc	Punctuation	.
13	sb	Symbol	X
14	tn	Number text letter	NUM
15	v	Verb	VERB

Table A.5. Mapping scheme between POS tags and universal POS tags of English language.

No.	POS Tag [39]	Description	Corresponding UPOS Tag [40]
1	DT	Determiner	DET
2	CD	Cardinal number	NUM
3	NN	Noun, singular	NOUN
4	NNS	Noun, plural	NOUN
5	NNP	Proper noun, singular	NOUN
6	NNPS	Proper noun, plural	NOUN
7	EX	Existential there, such as in the sentence There was a party.	PRON
8	FW	Foreign Word	X
9	PRP	Personal pronoun (PP)	PRON
10	PRP$	Possessive pronoun (PP$)	PRON
11	POS	Possessive ending	ADP
12	PDT	Predeterminer	PRON
13	RBS	Adverb, superlative	ADV
14	RBR	Adverb, comparative	ADV
15	RB	Adverb	ADV

Table A.5 (*continued*)

No.	POS Tag [39]	Description	Corresponding UPOS Tag [40]
16	JJS	Adjective, superlative	ADJ
17	LS	List item marker	.
18	JJR	Adjective, comparative	ADJ
19	JJ	Adjective	ADJ
20	MD	Modal	VERB
21	VB	Verb, base form	VERB
22	VBP	Verb, present tense, other than third person singular	VERB
23	VBZ	Verb, present tense, third person singular	VERB
24	VBD	Verb, past tense	VERB
25	VBN	Verb, past participle	VERB
26	VBG	Verb, gerund or present participle	VERB
27	WDT	Wh-determiner, such as which in the sentence Which book do you like better	PRON
28	WP	Wh-pronoun, such as which and that when they are used as relative pronouns	PRON
29	WP$	Possessive wh-pronoun, such as whose	PRON
30	WRB	Wh-adverb, such as when in the sentence I like it when you make dinner for me	ADV
31	TO	The preposition to	ADP
32	IN	Preposition or subordinating conjunction	ADP
33	CC	Coordinating conjunction	CONJ
34	UH	Interjection	X
35	RP	Particle	ADP
36	SYM	Symbol	X
37	#	Pound sign	X
38	$	Dollar sign	X
39	“	Opening quotation marks	.
40	”	Double or single quotation marks	.
41	(	Opening parenthesis, bracket, angle bracket, or brace	.
42	)	Closing parenthesis, bracket, angle bracket, or brace	.
43	,	Comma	.
44	.	End of sentence punctuation (. !?)	.
45	:	Mid-sentence punctuation (– ; : ... —)	.

References

[1] N. Kalchbrenner, P. Blunsom, Recurrent continuous translation models, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle, Washington, USA, 18–21 October 2013, 2013, pp. 1700–1709, URL https://aclanthology.org/D13-1176.

[2] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, in: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS’14, MIT Press, Cambridge, MA, USA, 8–13 December 2014, 2014, pp. 3104–3112.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17, Curran Associates Inc., Red Hook, NY, USA, 4–9 December 2017, 2017, pp. 6000–6010.

[4] J. Su, J. Zeng, D. Xiong, Y. Liu, M. Wang, J. Xie, A hierarchy-to-sequence attentional neural machine translation model, IEEE/ACM Trans. Audio Speech Lang. Process. 26 (3) (2018) 623–632.

[5] J. Zhang, M. Wang, Q. Liu, J. Zhou, Incorporating word reordering knowledge into attention-based neural machine translation, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Volume 1, Long Papers, Association for Computational Linguistics, Vancouver, Canada, 30 July–4 August 2017, 2017, pp. 1524–1534, URL https://aclanthology.org/P17-1140.

[6] S. Feng, S. Liu, N. Yang, M. Li, M. Zhou, K.Q. Zhu, Improving attention modeling with implicit distortion and fertility for machine translation, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan, 11–17 December 2016, 2016, pp. 3082–3092, URL https://aclanthology.org/C16-1290.

[7] T. Cohn, C.D.V. Hoang, E. Vymolova, K. Yao, C. Dyer, G. Haffari, Incorporating structural alignment biases into an attentional neural translation model, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 12–17 June 2016, 2016, pp. 876–885, URL https://aclanthology.org/N16-1102.

[8] Y. Cheng, S. Shen, Z. He, W. He, H. Wu, M. Sun, Y. Liu, Agreement-based joint training for bidirectional attention-based neural machine translation, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI’16, AAAI Press, Palo Alto, California, USA, 9–15 July 2016, 2016, pp. 2761–2767.

[9] T. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 17–21 September 2015, 2015, pp. 1412–1421, URL https://aclanthology.org/D15-1166.

[10] H. Chen, S. Huang, D. Chiang, J. Chen, Improved neural machine translation with a syntax-aware encoder and decoder, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Volume 1, Long Papers, Association for Computational Linguistics, Vancouver, Canada, 30 July–4 August 2017, 2017, pp. 1936–1945, URL https://aclanthology.org/P17-1177.

[11] K. Chen, R. Wang, M. Utiyama, L. Liu, A. Tamura, E. Sumita, T. Zhao, Neural machine translation with source dependency representation, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 7–11 September 2017, 2017, pp. 2846–2852, URL https://aclanthology.org/D17-1304.

[12] K. Chen, R. Wang, M. Utiyama, E. Sumita, T. Zhao, Syntax-directed attention for neural machine translation, Proc. AAAI Conf. Artif. Intell. 32 (1) (2018) 4792–4799, https://ojs.aaai.org/index.php/AAAI/article/view/11910.

[13] J. Li, D. Xiong, Z. Tu, M. Zhu, M. Zhang, G. Zhou, Modeling source syntax for neural machine translation, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Volume 1, Long Papers, Association for Computational Linguistics, Vancouver, Canada, 30 July–4 August 2017, 2017, pp. 688–697, URL https://aclanthology.org/P17-1064.

[14] S. Wu, D. Zhang, N. Yang, M. Li, M. Zhou, Sequence-to-dependency neural machine translation, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Volume 1, Long Papers, Association for Computational Linguistics, Vancouver, Canada, 30 July–4 August 2017, 2017, pp. 698–707, URL https://aclanthology.org/P17-1065.

[15] A. Eriguchi, K. Hashimoto, Y. Tsuruoka, Tree-to-sequence attentional neural machine translation, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Volume 1, Long Papers, Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016, 2016, pp. 823–833, URL https://aclanthology.org/P16-1078.

[16] R. Sennrich, B. Haddow, Linguistic input features improve neural machine translation, in: Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers, Association for Computational Linguistics, Berlin, Germany, 11–12 August 2016, 2016, pp. 83–91, URL https://aclanthology.org/W16-2209.

[17] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst 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, Association for Computational Linguistics, Prague, Czech Republic, 25–27 June 2007, 2007, pp. 177–180, URL https://aclanthology.org/P07-2045.

[18] P. Koehn, H. Hoang, Factored translation models, in: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Association for Computational Linguistics, Prague, Czech Republic, 29–30 June 2007, 2007, pp. 868–876, URL https://aclanthology.org/D07-1091.

[19] Y.K. Thu, A. Finch, A. Tamura, E. Sumita, Y. Sagisaka, Factored machine translation for Myanmar to English, Japanese and vice versa, in: Proceedings of the 12th International Conference on Computer Applications, ICCA 2014, Yangon, Myanmar, 17–18 February 2014, 2014, pp. 171–177.

[20] H. de Medeiros Caseli, I.A. Nunes, Factored translation between Brazilian Portuguese and English, in: A.C. da Rocha Costa, R.M. Vicari, F. Tonidandel (Eds.), Advances in Artificial Intelligence – SBIA 2010, Springer Berlin Heidelberg, Berlin, Heidelberg, 23–28 October 2010, 2010, pp. 163–172.

[21] M. García-Martínez, L. Barrault, F. Bougares, Factored neural machine translation architectures, in: International Workshop on Spoken Language Translation, IWSLT'16, 8–9 December 2016, 2016, pp. 8–9.

[22] Y. Wu, X. Lu, H. Yamamoto, S. Matsuda, C. Hori, H. Kashioka, Factored language model based on recurrent neural network, in: Proceedings of COLING 2012, The COLING 2012 Organizing Committee, Mumbai, India, 8–15 December 2012, 2012, pp. 2835–2850, URL https://aclanthology.org/C12-1173.

[23] E. Cho, J. Niehues, T.-L. Ha, M. Sperber, M. Mediani, A.H. Waibel, Adaptation and combination of nmt systems: the kit translation systems for IWSLT 2016, in: In International Workshop on Spoken Language Translation, IWSLT'16, December 2016, pp. 8–9.

[24] Y. Yin, J. Su, H. Wen, J. Zeng, Y. Liu, Y. Chen, Pos tag-enhanced coarse-to-fine attention for neural machine translation, ACM Trans. Asian Low-Resour. Lang. Inf. Process. 18 (4) (Apr. 2019).

[25] J. Niehues, E. Cho, Exploiting linguistic resources for neural machine translation using multi-task learning, in: Proceedings of the Second Conference on Machine Translation, Association for Computational Linguistics, Copenhagen, Denmark, 7–11 September 2017, 2017, pp. 80–89, URL https://aclanthology.org/W17-4708.

[26] L.H.B. Nguyen, H.D. Minh, H. Wen, D. Ding, T.L. Manh, Improving neural machine translation with pos tags, ICIC Express Lett., Part B: Appl. 12 (1) (2021) 91–98.

[27] R. Perera, T. Fonseka, R. Naranpanawa, U. Thayasivam, Improving English to sinhala neural machine translation using part-of-speech tag, arXiv preprint arXiv:2202.08882, 2022.

[28] Y. Pan, X. Li, Y. Yang, R. Dong, Multi-source neural model for machine translation of agglutinative language, Future Internet 12 (6) (2020), URL https://www.mdpi.com/1999-5903/12/6/96.

[29] X. Feng, Z. Feng, W. Zhao, B. Qin, T. Liu, Enhanced neural machine translation by joint decoding with word and pos-tagging sequences, Mob. Netw. Appl. 25 (5) (2020) 1722–1728.

[30] M. Junczys-Dowmunt, R. Grundkiewicz, MS-UEdin submission to the WMT2018 APE shared task: dual-source transformer for automatic post-editing, in: Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Association for Computational Linguistics, Belgium, Brussels, 2018, pp. 822–826, URL https://aclanthology.org/W18-6467.

[31] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 27–30 June 2016, 2016, pp. 770–778.

[32] J.L. Ba, J.R. Kiros, G.E. Hinton, Layer normalization, arXiv:1607.06450, 2016.

[33] B. Prachya, S. Thepchai, Technical report for the network-based asean language translation public service project, online materials of network-based asean languages translation public service for members, Tech. Rep., 2013.

[34] Z.Z. Hlaing, Y.K. Thu, M.M.N. Wai, S. Thepchai, N. Ponrudee, Myanmar pos resource extension effects on automatic tagging methods, in: In 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing, iSAINLP, 18–20 November 2020, 2020, pp. 1–6.

[35] D.Q. Nguyen, D.Q. Nguyen, D.D. Pham, S.B.a. Pham, A robust transformation-based learning approach using ripple down rules for part-of-speech tagging, AI Commun. 29 (3) (2016) 409–422.

[36] K.W.W. Htike, Y.K. Thu, Z. Zhang, W.P. Pa, Y. Sagisaka, N. Iwahashi, Comparison of six pos tagging methods on 10k sentences Myanmar language (Burmese) pos tagged corpus, in: 18th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing, 17–23 April 2017, 2017, pp. 17–23.

[37] B. Prachya, L. Vorapon, P. Sitthaa, K. Kanyanat, L. Dhanon, P. Charun, B. Monthika, K. Krit, S. Thepchai, The annotation guideline of LST20 corpus, CoRR, arXiv:2008.05055 [abs], 2020, arXiv:2008.05055, URL https://arxiv.org/abs/2008.05055.

[38] V. Sornlertlamvanich, N. Takahashi, H. Isahara, Building a Thai part-of-speech tagged corpus (ORCHID), J. Accoust. Soc. Jpn. (E) 20 (3) (1999) 189–198.

[39] M.P. Marcus, B. Santorini, M.A. Marcinkiewicz, Building a large annotated corpus of English: the Penn Treebank, Comput. Linguist. 19 (2) (1993) 313–330, URL https://aclanthology.org/J93-2004.

[40] S. Petrov, D. Das, R. McDonald, A universal part-of-speech tagset, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 23–25 May 2012, 2012, pp. 2089–2096, URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf.

[41] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A.F. Aji, N. Bogoychev, A.F.T. Martins, A. Birch Marian, Fast neural machine translation in C++, in: Proceedings of ACL 2018, System Demonstrations, Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018, 2018, pp. 116–121, URL https://aclanthology.org/P18-4020.

[42] D.P. Kingma, J. Ba Adam, A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015, URL http://arxiv.org/abs/1412.6980.

[43] G. Neubig, Z.-Y. Dou, J. Hu, P. Michel, D. Pruthi, X. Wang, Compare-mt: a tool for holistic comparison of language generation systems, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Association for Computational Linguistics, Minneapolis, Minnesota, 2–7 June 2019, 2019, pp. 35–41, URL https://aclanthology.org/N19-4007.

[44] W. Xu, M. Carpuat, EDITOR: an edit-based transformer with repositioning for neural machine translation with soft lexical constraints, Trans. Assoc. Comput. Linguist. 9 (2021) 311–328, URL https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00368/1923848/tacl_a_00368.pdf.

[45] J. Gu, C. Wang, J.Z. Junbo, Levenshtein Transformer, Curran Associates Inc., Red Hook, NY, USA, 2019.

[46] M.E. San, Y.K. Thu, Z.Z. Hlaing, H.M. Nwe, T. Supnithi, S. Usanavasin, A study of levenshtein transformer and editor transformer models for under-resourced languages, in: 2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing, iSAI-NLP, 2021, pp. 1–6.

[47] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 7–12 July 2002, 2002, pp. 311–318, URL https://aclanthology.org/P02-1040.

[48] M. Popović, chrF: character n-Gram F-score for automatic MT evaluation, in: Proceedings of the Tenth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 392–395, URL https://aclanthology.org/W15-3049.

[49] W. Fisher, J. Fiscus, Better alignment procedures for speech recognition evaluation, in: 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, 1993, pp. 59–62.