SN Computer Science (2023) 4:841
https://doi.org/10.1007/s42979-023-02309-8

ORIGINAL RESEARCH

Parts of Speech Tagged Phrase-Based Statistical Machine Translation System for English → Mizo Language

Chanambam Sveta Devi[1] · Amit Kumar Roy[1] · Bipul Syam Purkayastha[1]

Received: 27 March 2023 / Accepted: 2 September 2023

Abstract
In Natural Language Processing (NLP), Parts of Speech (PoS) tagging is one of the most crucial steps of pre-processing in every language system. PoS tagging is the process of selecting the best suitable 'part of speech category' or 'lexical class label' for each token in a phrase in a natural language. It is usually the initial stage of an NLP task like machine translation, with further stages, including chunking, parsing, etc. The key objective of our work is to create an English-to-Mizo Machine Translation (MT) system using PoS tag corpus. Here, we use factored Phrase-Based Statistical MT (F-SMT) to address the issue of text translation from English into Mizo, one of the under-resource language pairs. During the process, we discovered that better accuracy of the system can be achieved by combining the training with PoS-featured data. Experimental results achieved by employing automatic evaluation metrics demonstrated that our F-SMT with PoS tagging model outperformed the baseline phrase-based model and other factored models when applying the various n-gram parameters. Our F-SMT with PoS tagging exhibits an increase in the automated scores of translation; it achieved scores of 14.27 (BLEU), 47.70 (*F*-measure), 30.30 (METEOR), 37.60 (Precision), 52.40 (Recall), and 75.30 (TER).

Keywords Factored model · PoS tagging · Mizo language · BLEU

Introduction

Natural Language Processing is a subfield of Artificial Intelligence (AI) that works with human natural languages. It is a method of processing, exploring, examining, and comprehending vast amounts of text data with machines. It provides a system that assists a machine or computer in understanding, interpreting, and processing human language, as well as in resolving ambiguity in various natural languages. PoS tagging is an important pre-processing module for any type of NLP task. It is the method of identifying and labeling up to the token of a given text using proper PoS components, i.e., noun, pronoun, adjective, or lexical class maker. As a result, it provides details regarding the usage of a word in a sentence or context other than that of a general. This extra data is helpful when used in pre- or post-processing approaches of various NLP applications.

The state-of-the-art method of SMT, known as phrase-based SMT (PB-SMT) models, is confined to the mapping of short text chunks with no explicit use of linguistic information, whether morphological, syntactic, or semantic. This model can perform better when incorporated with additional linguistic information. The translation model should include language information more closely for two reasons, though [1]: (i) Translation models that use more language information, like lemmas, rather than surface forms (words), may get access to richer statistics and overcome data sparsity problems caused by limited training data. (ii) The best way to describe many aspects of translation is on a morphological, syntactic, or semantic level. Direct modeling of these traits is made possible by the translation model's access to this information. For instance, while local agreement constraints are

This article is part of the topical collection "SWOT to AI-embraced Communication Systems (SWOT-AI)" guest edited by Somnath Mukhopadhyay, Debashis De, Sunita Sarkar and Celia Shahnaz.

✉ Amit Kumar Roy
amitroy.cs@gmail.com

Chanambam Sveta Devi
chsveta91@gmail.com

Bipul Syam Purkayastha
bipul_sh@hotmail.com

[1] Department of Computer Science, Assam University, Silchar 788011, Assam, India

Published online: 03 November 2023

present in morphology, general syntactic principles mostly regulate reordering at the sentence level.

A Factored translation model [1], an extension to PB-SMT, is the best example for adding any additional information, including morphological, at decoding time. PB-SMT's fundamental problem is that it translates phrases of sentences without explicitly using linguistic annotation, even with the fact that this would seem to be beneficial for a smooth translation outcome. In this structure, a word is not just a token but also a collection of factors, since a set of tags augments every word in Factored models. For instance, a basic word may be represented as a group of surface form (words) along with the lemma, PoS tag, word class, and morphological information.

It is clear that new representation is more complex than the surface form of the word. Because factored models focus on word-level enrichments, they certainly address the morphological challenge, which matches the current situation. In factored models, translation works are usually divided into two translation phases and one generation phase. The first is the translation of a source text into a target lemma. The lemma and other elements are used to construct the final form after morphological and PoS components are translated into target forms in the second part. Factored models are built in a similar manner as phrase-based systems are. The translation step in these models operates at a phrase level, while the generating phases are word-level operations.

The process for converting English text into Mizo text in the F-SMT is explained step by step:

Factored representation: (surface form: keini), (lemma: kei), (PoS: PRP), (count: Plural), (case: nominative).

Translation (mapping lemmas): kei → I|we|see|bird.

Translation (mapping morphology): PRP| plural-nominative-pronoun → PRP|plural.

PRP|Singluar.

Generation (generating surface forms):

— I|PRP|plural — we.

— I|PRP|singular — I.

— we|PRP|plural — we.

Machine Translation is a difficult process, as different natural languages with their linguistic distinctions add more difficulties for it. One such language pair is English–Mizo. We highlight some distinctions between the English language and the Mizo language. English language has basic morphology with a Subject–Verb–Object (S–V–O) sentence structure and is non-tonal. The most prevalent method of word generation in English is derivation, such as “Im + possible” and “Un + kind”. On the other hand, Mizo is an agglutinative language with a rich morphological structure with an Object–Subject–Verb (O–S–V) structure; however, it follows English as S–V–O. The Mizo language is a tonal language because tone dictates the lexical meaning of words [2]. There are a total of eight tones in Mizo, including four long tones and four short tones. In the Mizo language's tonal words, the use of diacritics is not explicitly stated. It can be challenging to assign PoS at times since the context in which a word is used might change how it is understood. A lexical root is followed by one or more affixes in Mizo words. Person, number, gender, and case markers inflect Mizo words.

Until now, there have been only a few scientific articles on machine translation from English to Mizo. Therefore, the baseline model of this language translation is compared to the baseline of PoS tag translation to gain a better understanding. The uniqueness of this research is in obtaining early findings of the English to Mizo PB-SMT system. This phrase-based approach was chosen as it is easier to execute and does not require any linguistic information. Furthermore, using the rule-based method, we will not be able to use any readily available Mizo language resources or tools. Recently, Google Translate[1] added the Mizo translation to their system, that uses Neural Machine Translation (NMT) and dictionary-based translation, but the result of the translations is still very imprecise.

This paper implements the factored SMT model using the PoS-tagged dataset of the English–Mizo language pair and compares it to the PB-SMT baseline model. The contribution of this paper includes.

(i) Manual development of a PoS-tagged dataset of the Mizo corpus of the NPLT (National Platform of Language Technology) domain using tag set of Pakray’s research [3], consisting of 24 tags and other new tags.

(ii) Implementation of baseline PB-SMT model and F-SMT model. The F-SMT model contains two different parameters, viz FPoST-1 and FPoST-2.

(iii) Comparison of the results of these systems with different evaluation metrics and explanation of how our design enhances translation outcomes.

The expected translation outcome was evaluated using automatic evaluation tools like BLEU (Bilingual Evaluation Under-study) [4], TER (Translation Error Rate) [5], METEOR [6], F-measure, Precision, and Recall.

The following section address about the factored translation model using only the PoS tag with the Mizo language characteristics, as well as some related research findings on translation and their existing works, with the brief details on the F-SMT model, which is followed by an experimental set of results and discussion, the conclusion, and further work plan.

Factored PB-SMT

The Factored SMT model is an expansion of the PB-SMT model, the most used SMT technique. The phrase model

[1] https://translate.google.com/.

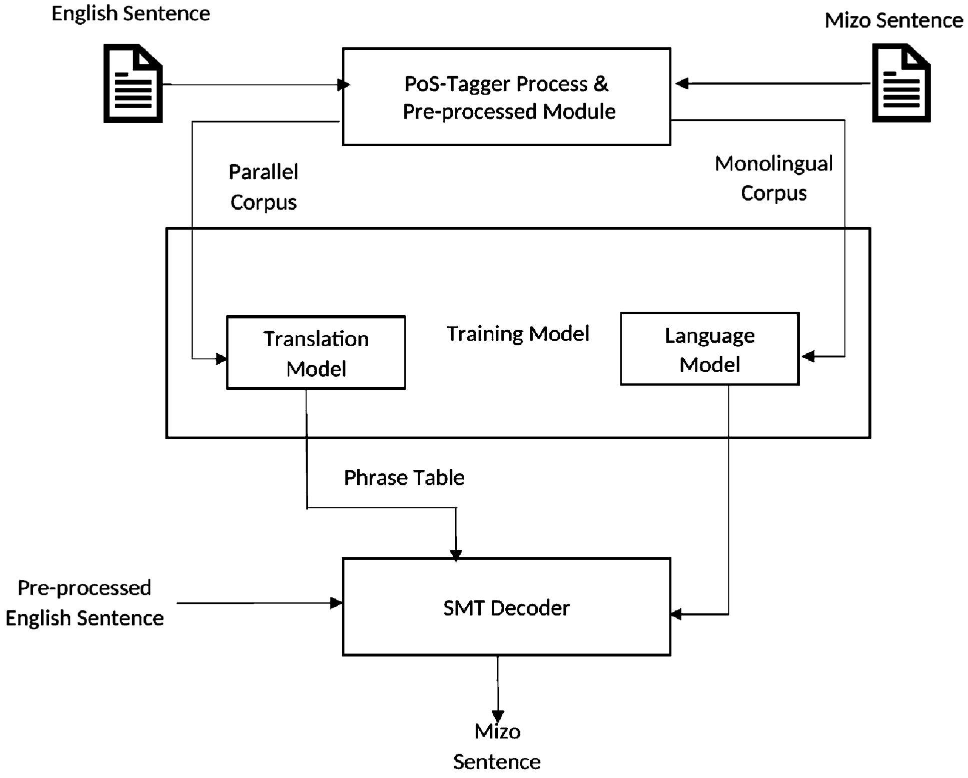

Fig. 1 The architecture of the F-SMT model

translates short text passages and phrases without taking into account any linguistic details. The translation probability for changing a source sentence (s) into a target sentence (t) is given by the Bayes theorem [7].

$$P\langle t \mid s\rangle = P\langle s \mid t\rangle P(t)/P(s) \tag{1}$$

$$t = \arg\max P\langle s \mid t\rangle P(t)/P(s) \tag{2}$$

$$t = \arg\max P\langle s \mid t\rangle P(t). \tag{3}$$

As the probability of the source text $P(s)$ is constant, the denominator $P(s)$ is removed from Eq. (3). The translation model provides $P(s/t)$, whereas the language model provides $P(t)$. In addition, a decoder is needed to identify the best translation, which, given a source sentence s and a target sentence t, generates the best possible translation or, alternatively, an n-best list of the most likely translations. The most likely translation (t) out of all possible target language sentences is chosen using *argmax* with the probability of translation and language model to calculate the probability of the best translation (t). A SMT technique called PB-SMT uses a phrase rather than as a single word as a translation unit. When using PB-SMT, the input text is divided into a certain number of phrases, denoted as i. Each of these phrases, t_i, in the source language is then translated into a corresponding target phrase, e_i. The following equation, which includes a phrase reordering model, is used to calculate the translation of sentence f [8].

$$P\left\langle f_{i=1}^{-1} \mid s_{i=1}^{-1} \right\rangle = \prod_{i=1}^{1} \phi\left\langle f_i \mid s_i \right\rangle d\left(start_i - end_{i-1} - 1\right).$$

If $\phi\left\langle f_i \mid s_i \right\rangle$ represents the phrase translation probability and $d\left(end_{i-1} - 1\right)$ represents the distance-based reordering model, and $start_i$ and end_i are concerned as the initial and last words of the source sentence that translated into the target phrase i. The F-SMT process is shown in Fig. 1 given below.

The source sentences and their corresponding translated target sentences are prepared by the parallel corpus collection. If no parallel corpus is available, we must prepare it manually or collect it from trusted secondary sources. After collecting the parallel corpus, pre-processing steps were performed using Tokenization, True-casing, and cleaning [9]. Tokenization is obtaining a word or a piece of punctuation by leaving a space between them. Whereas True-casing includes using the most frequent case for the first letter of each phrase to minimize data sparsity. While cleaning is the process of removing an empty sentence from the corpus of texts. Furthermore, cleaning also limits sentence length because large sentences have a higher chance of erroneous translation than short sentences [9].

In the next step, the training of language model is trained using the monolingual corpus of the target sentences, while the training of translation model uses the parallel corpus. When a translation model is trained, a phrase table is produced that includes phrases from the corpus and their likelihood of recurrence. While the language model training generates n-gram sentences in the monolingual target language. The input text is decrypted and translated as a consequence of employing translation and language models.

Previous Works

Due to the effectiveness of utilizing extra language features in multiple NLP assignments, multiple methods have been suggested for integrating supplementary linguistic information into statistical phrase-based systems. Our study involves the integration of supplementary linguistic factors like PoS tags into PB-SMT models, with the purpose of enhancing the quality of translation for language pairs that have limited linguistic resources. Our research has a connection to the following studies.

The first machine translation system to use linguistic information was statistical machine translation (SMT). The Moses toolkit [10] was used to help implement the source and target factors. They were used for PoS tags, morphological tags, "surface" forms, "lemmas", and other tag combinations [1]. According to [4], the suggested method will improve bilingual evaluation understudy (BLEU) results over the existing phrase-based SMT technique by up to 2%. SMT system have been successful at generating factors and factored SMT model that incorporated linguistic aspects to enhance translation quality and address grammatical error, data sparsity, and fluency for morphologically diverse language. They explore an early method of translating from French to English that made use of linguistic information in their work [11]. Another implementation report [12] discusses using factored models for English–Latvian and English–Lithuanian SMT systems. The languages of Latvia and Lithuania are very inflectional. They are morphologically complex, have a flexible phrase structure, and are quite ambiguous, which makes translation data scarce. By separating each token into its stem and suffix components and considering them as independent models for Lithuanian–English machine translation, they identified an approach to this problem. While morphological tags are employed for English–Latvian as an extra language model along with suffixes. Through human evaluation, their work asserts a considerable advancement over baseline SMT. Through human inspection, their work claims a notable advancement over the initial SMT. The advantages of employing several parameters were illustrated by the same testing phrase-based MT for English–Czech [13]. This paper uses a variety of models that take into consideration variables including word form, lemma, and morphological tags. The BLEU score report, which acts as the conclusion of his work, shows how multifactor SMT routinely beats baseline SMT. Another study uses fixed-length word suffixes that, in some respects, resemble PoS tags to create a factored SMT model [14]. Their approach minimizes the complexity of the language model and improves the outcomes' grammatical accuracy. Their research demonstrates an improvement over the SMT's baseline.

Translations into them are limited by the poor output translation quality, in contrast to translations from morphologically complex languages. The problem of data sparseness is a significant concern for morphologically complex languages. There are numerous publications on data sparsity for morphologically rich languages like Latvian, Lithuanian, Croatian, Tamil, Malayalam, Mizo, Hindi, Kannada, and Farsi. In their study, they address the issue of data sparsity while translating a morphologically complex language [15]. They recommended a method that develops hidden morphological forms and feeds them into the training corpora. Through experiments involving the translation of English into Hindi and Marathi, their suggested solution is said to improve the quality of translation.

In addition to factors on the corpus, pre-processing works are also provided. There is a factoring SMT system for translating from English to Tamil [16]. Lemma, PoS, and combine tags are factors in their model within the source side, whereas lemma, PoS tags, and morphological data are factors on the target side. Here they create an innovative technique for data pre-processing for the English source language into Tamil target language. And the pre-processed sentences are used in the training with a factored SMT model. Finally, using the components produced by the SMT model, Tamil morphological generators create words in their surface form. The output result performs better than other systems, including Google Translate. Another similar endeavor is pre-processing to modify the structure of the input text (English) by adding PoS tags [17]. To improve the English sentences more comparable to the more complex Spanish and Catalan target sentences, they use PoS tags with them.

The study by [18] examines the factored SMT model in comparison to the standard SMT system method for translating the morphologically rich Kannada language. For the factored model, they develop language models based on surface form and PoS tags. According to their report, the factored model offers a 25% increase in BLEU over the baseline model.

This is the existing work on PoS tagging research on the Mizo language. Only a few works are done in the study of Part of Speech in this language. The authors [3] proposed the framework for the development of the Mizo PoS system by proposing 24 tagsets for the Mizo language using the Penn Treebank tagset system. In these works, they also built a Mizo to English dictionary, which comprises English phrases and their Mizo meanings, synonyms for that word, and the PoS tag for each synonym. A combination of automatic and manual procedures was used to create the dictionary. It has 26,407 entries. The paper has been developed by [2]: An annotated preliminary study, discussed the unique qualities of the Mizo language as well as the Mizo tagging system's limitations; and proposed a tag list of 37 tagsets and defines the grammatical information of tokens in a text. Attempts were made to investigate various tagging systems, which aid in the testing of the Mizo language's detained morphology. PoS tagging for Mizo Language using a CRF [19] reported that Mizo language development has been limited due to a lack of resources, which employs the conditional random field stochastic model (CRF). The CRF, a type of probabilistic classifier, considers both a word's context and the likelihood of transitions between tags in the training data. To evaluate the system, a collection of around 30,000 words was gathered and manually labeled with the suggested set of tags. Across different training and testing sets, the tagger performed with 89.46%—accuracy, 89.3%—$F1$-score, 89.42%—precision, and 89.48%—recall. And with the few articles of machine translation from English to Mizo language [20], the NMT system was educated for translating English to Mizo, using a parallel corpus of 10,675 sentences. Its effectiveness was tested on a separate dataset of 100 sentences, and the results showed that the system was satisfactory in terms of fluency, but not accuracy. Ref. [21] study was to evaluate the performance of the same NMT system in various domains using multiple test datasets. In a study on English to Mizo machine translation systems, Ref. [22] utilized a training dataset sourced from different online platforms to compare the effectiveness of SMT and NMT systems. Ref. [23] proposed an experimental test on the English–Mizo statistical machine translation with the Bible corpus. The system was analyzed using the automatic scoring methodologies BLEU and METEOR score, as well as manually evaluated by linguistic experts. SMT systems with BLEU score of 18.71 for English to Mizo and score of 19.44 for Mizo to English perform better than other MT systems when trained with the Language Model's 5-g order. The outcomes of the automatic evaluation demonstrate that the MT system performs better as the n-gram order of the LM increases. To get better translation outcomes in PB-SMT, we choose to use manually labeled PoS tags created with the Mizo resources to produce PoS, lemma, and other linguistic characteristics.

In this paper, we report our testing of PoS tag datasets of English–Mizo on the phrase-based statistical machine translation. According to our knowledge, currently, we had to manually construct our PoS tagger due to the lack of a tagged dataset for this language. This is the first English to Mizo PoS tag dataset that has been applied to the PB-SMT. The PoS tag models were compared with the baseline model of the PB-SMT.

Table 1 Statistics description of the baseline datasets

Types	No. of sentences	No. of tokens
Train (English)	8250	161,379
Train (Mizo)		207,626
Tune (English)	1000	19,777
Tune (Mizo)		23,432
Test (English)	750	15,248
Test (Mizo)		18,413

Experimental Setup

There are significant language differences between English and Mizo, as already mentioned. The translation is made more difficult by their morphological and structural variations. We use unique pre-processing tools for both English and Mizo phrases, then provide training to handle this. More information is provided in the subsections below.

Data Collection and Pre-processing

As there is no existing parallel corpus available for English to Mizo, we utilize a small parallel corpus that is constructed by manually translating through a linguistic person after collecting the monolingual English text from the National Platform of Language Technology[2] (NPLT). The parallel corpus should be clean of typos and inconsistencies.

We have 10,000 sentences of the parallel corpus, which are arranged for the training, tuning, and testing process using Sklearn code. Both the systems need to be pre-processed, in the pre-processing methods, to reduce noise, non-ASCII special characters are eliminated from the parallel corpus. After cleaning the corpus, MOSES tokenizer is used to tokenize it [10]. Table 1 gives the number of sentences and tokens for the given corpus.

System Training

The primary idea of this study is to establish which translation model based on factored produces better results and also to identify any problems that arise while translating texts from English to the Mizo language. While training the language model, we use KENLM [24] that applies the

[2] https://nplt.in/demo/.

Table 2 Statistical result of the systems shows the BLEU, *F*-measure, METEOR, Precision, Recall, and TER scores of the testing data set

Models	BLEU	*F*-measure	METEOR	Precision	Recall	TER
Baseline	12.66	39.40	20.90	38.80	39.70	85.80
FPoST-2	13.16	44.00	26.90	34.00	48.80	77.70
FPoST-1	14.27	47.70	30.30	37.60	52.40	75.30

n-gram approach, the 5-g approach is used for generating the baseline LM (surface.lm), while SRILM [25] of 7-g approach is used for factored LM (pos.lm). The word alignment and reordering is done using the GIZA + + [26] tool. The training of the translation model incorporates both "surface.lm" and "pos.lm" files. While decoding the output of both trainings helps to determine the best suitable translation. We employed the automatic evaluation metrics BLEU, METEOR, and TER to evaluate the result of the decoding task [8].

We are using two systems of the test set, one with the baseline PB-SMT for the "surface" (word) form and the other factored model used PoS tag along with the surface form. For comparing the system, we tested on different models according to their parameter settings on the same dataset. The first model is the PB-SMT which is the baseline PB-SMT without any linguistic annotation feature applied to the dataset. Again we applied two different models on the factored-PoS tag on the same data. FPoST-1 "PoS tag was added to the Mizo side" and FPoST-2 "PoS tag was added to both of the English and Mizo".

The corpus used in the baseline PB-SMT system was carefully translated by a linguistic person. In addition, the tag sets of the two languages are different. Due to each difference in language characteristics, we must use two separate tag sets for PoS tagging. In terms of English, we used the Penn Treebank tag set NLTK tagged [27]. While in Mizo, we followed the PoS tag set of Pakray's research [3], consisting of 24 tags and other new tags as there are no available NLP tools for the Mizo language. The same reference text used in the first system is used to evaluate the translation output of the factored model.

Example 1. PoS Tag Dataset

English tag: monitoring|VBG the|DT epidemic|JJ and|CC prevention|NN policies| NNS.|.

Manual tag: epidemic|NNS te|PRP vil|VB that|RB leh|CC invenna|NN policies| NNS te|PRP.|SYM.

Result and Discussion

This part shows the results of the experiment from the parallel corpora analyzed using automated evaluation measures. In addition, we looked at how well the PB-SMT and F-SMT systems predicted translations.

Results for BLEU Score

It is known that using more words in training increases the BLEU score based on test results using only the "surface" (word) form. A low baseline PB-SMT BLEU score can be caused by a number of factors, including a lack of vocabulary, words that are difficult to understand (ambiguous words), and the alignment discrepancy between text instruction in English and Mizo.

Even though the data set has more OOV, still BLEU score is fine because many of the OOV is untranslated name entities and foreign words, as well as similar reference sentences.

The baseline model, like OOV, cannot translate ambiguous words. The incorrect translation of an ambiguous word is caused by a lack of training corpus, the phrase table does not contain all translation options for such ambiguous words.

We created three different models of translations. The PB-SMT baseline system is designed with default settings and is referred to as Baseline. Furthermore, the PoS tag data set was applied to the same train data with two different parameters models denoted as FPoST-1 and FPoST-2. The evaluation of these 3 models is performed on a test set of 750 lines.

According to Table 2, model FPoST-1 has the better BLEU score and is closely followed by model FPoST-2. However, we find that a few English words remain untranslated. The baseline model outperforms the previous two, but this could be due to the small size of the dataset and vocabulary.

Results for METEOR and TER Score

METEOR finds the exact, stemmed, synonymic, and paraphrase matched among the desired translations with the reference translations. Table 2 displays the test sets' METEOR, *F*-measure, and TER scores.

The fact that only an exact match is feasible in the alignment of the unigrams in the target Mizo language is generally blamed for METEOR's lower result of English–Mizo translation.

The Mizo language does not yet have any established stemming or synonyms for the target language. The penalty increases when unigrams are grouped into chunks, which lowers the overall score. Furthermore, the outcome could differ if only one reference translation is used.

F-measure only takes into account unigrams that provide a larger percentage of matches since it is the harmonic mean of accuracy and recall. In comparison to other models, PB-SMT obtains the greatest METEOR and *F*-measure ratings.

Translation error rate, which is the error rate or the minimal insertion, deletion, and replacement activities required to convert the resultant translation to a single reference translation are the two terms that describe this process. The need for less post-editing work is, therefore, indicated by a low TER score. The FPoST-1 model has the lowest TER score of all the models with a score of 75.30. By comparing the error rate of desired translations to a single reference translation, the models achieve excellent TER ratings.

In Example 2, the model FPoST-1 and FPoST-2 have best BLEU score than the baseline model which decrease the BLEU score point to 1.28 and 2.31 point from the model FPoST-1 and model FPoST-2 and the model FPoST-1 increases the BLEU score by 1.03 points from the model FPoST-2. The main issues in the low scores are the OOV's not translating a phrase and the incorrect translation of a phrase, whether it be for an unclear word or another. These models experienced problems due to a limited amount of parallel corpora. This only detected a very low percentage of the overall ambiguous word patterns reported in the phrase database. In addition, another issue in our testing resulted from an inaccurate word alignment, which happens frequently when a paraphrase translation is utilized in the parallel corpus.

Example 2. ***Sentences of Baseline SMT Model.***

English PoS-tag: whereas it is easy from the get-go for some, it can be challenging.

Reference tag: mi tam tak tan chuan awlsam te a pek theih a nih lai hian mi thenkhat tan chuan buaithlak tak thil a ni ve thei a ni.

Mizo PoS-tag: *industry a awlsam get—go a ni a, a biochemical thei a ni.*

BLEU score: 8.52.

Sentences of FPoST-1

English PoS-tag: whereas|IN it|PRP 's|VBZ easy|JJ from|IN the|DT get-go|NN for|IN some|DT,|, it|PRP can|MD be|VB challenging|VBG.|.

Reference tag: mi|PRP tam|JJ tak|JJ tan|JJ chuan|CC awlsam|JJ tea|PP pek|VB theih|VB a|PRP nih| VB lai|RB hian|PRP mi|PRP thenkhat|JJ tan|JJ chuan|CC buaithlak|JJ tak|JJ thil|PRP a|PRP ni|VB ve|RB thei|VB a|PRP ni|VB.|SYM.

Mizo PoS-tag: *tan|JJ chuan|CC buaithlak|JJ zawk|NN a|PRP awlsam|JJ tak|JJ a|PRP ni|VB.|SYM.*

BLEU score: ***10.83***

Sentences of FPoST-2

English PoS-tag: whereas|IN it|PRP 's|VBZ easy|JJ from|IN the|DT get-go|NN for|IN some|DT,|, it|PRP can|MD be|VB challenging|VBG.|.

Reference tag: mi|PRP tam|JJ tak|JJ tan|JJ chuan|CC awlsam|JJ tea|PP pek|VB theih|VB a|PRP nih|VB lai|RB hian|PRP mi|PRP thenkhat|JJ tan|JJ chuan|CC buaithlak|JJ tak|JJ thil|PRP a|PRP ni|VB ve|RB thei|VB a|PRP ni|VB.|SYM.

Mizo PoS-tag: *tan|JJ chuan|CC buaithlak|JJ zawk|NN a|PRP awlsam|JJ tak|JJ a|PRP ni|VB.|SYM.*

BLEU score: 9.80.

Conclusion

In this paper, we developed the PB-SMT model for the English–Mizo language utilizing two factored parameters, FPoST-1, and FPoST-2 and compared those results to the standard phrase-based SMT. The BLEU, METEOR, and TER scores were utilized to determine the expected translations. The translation model with FPoST-1 was found to be slightly more successful in Mizo than the baseline and FPoST-2. The results after adding linguistic information increased by 14.27 (BLEU), 47.70 (*F*-measure), 30.30 (METEOR), 37.60 (Precision), 52.40 (Recall), and 75.30 (TER).

We use a small dataset for our testing purposes and also manually tagged. If we have Mizo-specific NLP tools, we can automate the factoring process by combining other features to enhance our results even further. The objective is to develop methods that close the linguistic diverging gap because MT models never give the best word alignments for languages with far-off linguistic features. And we do not compare the results to those of Google Translate.

In the future, this research will try to look into English affixation and reduplication translated into Mizo while simultaneously resolving the noise issue in the dataset, we would also try to incorporate a post-editing technique for improving the quality of our system.

Data availability Data can be provided under special request.

Declarations

Conflict of Interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in the paper. The authors declare no conflict of interest.

References

1. Koehn P, Hoang H. Factored translation models. In: Proceedings of the 2007 Joint Conference on empirical methods in natural

language processing and computational natural language learning (EMNLP-CoNLL), 2007; pp. 868–876.

2. Nunsanga MV, Pakray P, Lalngaihtuaha M, Lolit Kumar Singh L. Part-of-speech tagging in Mizo language: a preliminary study. In: Data Intelligence and Cognitive Informatics: Proceedings of ICDICI 2020, 2021; pp. 625–635. Springer.
3. Pakray P, Pal A, Majumder G, Gelbukh A. Resource building and parts-of-speech (pos) tagging for the Mizo language. In: 2015 Fourteenth Mexican International Conference on Artificial Intelligence (MICAI), 2015; pp. 3–7 (2015).
4. Papineni K, Roukos S, Ward T, Zhu W-J. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002; pp. 311–318.
5. Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J. A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, 2006; pp. 223–231.
6. Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the Ninth Workshop on Statistical Machine Translation, 2014; pp. 376–380.
7. Stone JV. Bayes' rule: a tutorial introduction to Bayesian analysis. Sebtel Press; 2013. https://doi.org/10.13140/2.1.1371.6801.
8. Koehn P, Och FJ, Marcu D. Statistical phrase-based translation. In: Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 2003; pp. 127–133.
9. Devi CS, Purkayastha BS. Steps of pre-processing for English to Mizo smt system. In: Machine Learning, Image Processing, Network Security and Data Sciences: Second International Conference, MIND 2020, Silchar, India, July 30–31, 2020, Proceedings, Part II 2, 2020; pp. 156–167. Springer.
10. Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, et al. 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, 2007; pp. 177–180.
11. Brown PF, Della Pietra SA, Della Pietra VJ, Lafferty J, Mercer RL. Analysis, statistical transfer, and synthesis in machine translation. In: Proceedings of the Fourth Conference on theoretical and methodological issues in machine translation of natural languages, 1992.
12. Skadina I, Vasiljevs A. Human language technologies: the Baltic perspective: proceedings of the fourth International Conference, Baltic HLT 2010 vol. 219. Ios Press, 2010.
13. Bojar O. English-to-Czech factored machine translation. In: Proceedings of the Second Workshop on Statistical Machine Translation, 2007; pp. 232–239.
14. Razavian NS, Vogel S. Fixed length word suffix for factored statistical machine translation. In: Proceedings of the ACL 2010 Conference Short Papers, 2010; pp. 147–150.
15. Dungarwal PD. Reordering models for statistical machine translation: a literature survey. Bombay: Indian Institute of Technology; 2014.
16. Kumar MA, Dhanalakshmi V, Soman K, Rajendran S. Factored statistical machine translation system for English to Tamil language. Pertanika J Soc Sci Hum. 2014;22(4):1045–61.
17. Ueffing N, Ney H. Using PoS information for SMT into morphologically rich languages. In: 10th Conference of the European Chapter of the Association for Computational Linguistics; 2003.
18. Shivakumar K, Shivaraju N, Sreekanta V, Gupta D. Comparative study of factored smt with baseline smt for English to kannADa. In: 2016 International Conference on Inventive Computation Technologies (ICICT), 2016; vol. 1, pp. 1–6. IEEE.
19. Nunsanga MV, Pakray P, Lallawmsanga C, Singh L. Part-of-speech tagging for Mizo language using conditional random field. Computación y Sistemas. 2021;25(4):803–12.
20. Thihlum Z, Khenglawt V, Debnath, S. Machine translation of English language to Mizo language. In: 2020 IEEE International Conference on cloud computing in emerging markets (CCEM), 2020; pp. 92–97. IEEE.
21. Pathak A, Pakray P, Bentham J. English–mizo machine translation using neural and statistical approaches. Neural Comput Appl. 2019;31(11):7615–31.
22. Lalrempuii C, Soni B, Pakray P. An improved English-to-Mizo neural machine translation. Trans Asian Low-Resour Lang Inform Process. 2021;20(4):1–21.
23. Devi CS, Purkayastha BS, Meetei LS. An empirical study on English-Mizo statistical machine translation with bible corpus. Int J Electric Comput Eng Syst. 2022;13(9):759–65.
24. Heafield K. Kenlm: Faster and smaller language model queries. In: Proceedings of the Sixth Workshop on Statistical Machine Translation, 2011; pp. 187–197.
25. Stolcke A. SRILM an extensible language modeling toolkit. In: Seventh International Conference on spoken language processing. 2022.
26. Casacuberta F, Vidal E. Giza++: training of statistical translation models. 2007; Retrieved October 29, 2019.
27. Santorini B. Part-of-speech tagging guidelines for the penn treebank project (3rd revision). Technical Reports (CIS), 1990; 570.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.