Reordering of Source Side for a Factored English to Manipuri SMT System

Original Scientific Paper

Indika Maibam
Department of Computer Science
Indira Gandhi National Tribal University,
Kangpokpi, Imphal, Manipur, India
maibam.indika@igntu.ac.in

Bipul Syam Purkayastha
Department of Computer Science,
Assam University, Silchar, Assam, India
bipul_sh@hotmail.com

Abstract *– Similar languages with massive parallel corpora are readily implemented by large-scale systems using either Statistical Machine Translation (SMT) or Neural Machine Translation (NMT). Translations involving low-resource language pairs with linguistic divergence have always been a challenge. We consider one such pair, English-Manipuri, which shows linguistic divergence and belongs to the low resource category. For such language pairs, SMT gets better acclamation than NMT. However, SMT's more prominent phrase-based model uses groupings of surface word forms treated as phrases for translation. Therefore, without any linguistic knowledge, it fails to learn a proper mapping between the source and target language symbols. Our model adopts a factored model of SMT (FSMT3*) with a part-of-speech (POS) tag as a factor to incorporate linguistic information about the languages followed by hand-coded reordering. The reordering of source sentences makes them similar to the target language allowing better mapping between source and target symbols. The reordering also converts long-distance reordering problems to monotone reordering that SMT models can better handle, thereby reducing the load during decoding time. Additionally, we discover that adding a POS feature data enhances the system's precision. Experimental results using automatic evaluation metrics show that our model improved over phrase-based and other factored models using the lexicalised Moses reordering options. Our FSMT3* model shows an increase in the automatic scores of translation result over the factored model with lexicalised phrase reordering (FSMT2) by an amount of 11.05% (Bilingual Evaluation Understudy), 5.46% (F1), 9.35% (Precision), and 2.56% (Recall), respectively.*

Keywords*: Factored SMT, Reordering, Factoring, English, Manipuri, Automatic Evaluation*

1. INTRODUCTION

Machine Translation (MT) is defined as a "loop consisting of three steps in which, i) a source constituent is detected, ii) required information including syntactic, semantic and other types of information related to the constituent is collected, and iii) finally, it is transferred to a target form which is the end of the translation process for that constituent" [1]. By description, implementing an MT system seems straightforward and uncomplicated. Still, given the variety of languages spoken worldwide, each of which belongs to a separate family and has its unique linguistic structure, MT is not a simple procedure. The major issue is that these difficulties differ depending on the language combination under examination. Some of the challenges in our work are:

- Linguistic differences and their complexity
- Low resource category
- Unavailability of natural language processing (NLP) tools for Manipuri

MT is a very challenging task. The diversity of languages with linguistic differences between them, along with the inherently ambiguous nature, further amplifies the challenges. The language pair English-Manipuri is one such. We highlight a few linguistic differences between English and Manipuri. English has rudimental morphology with Subject-Verb-Object (S-V-O) sentences and non-tonal. The derivation is the most common process of word formation in English. For example, "*un+happy*", "*pre+judge*". In comparison, Manipuri has prolific morphology and agglutinating with Subject-Object-Verb (S-O-V) structure. Manipuri shows a variance from other Tibeto-Burman categories of languages in that it gives prominence to tense rather than mood.

Manipuri also shows tonal contrast, with two levels - high falling and level. For example, The word "*tummi*" can mean "*sleeping*" (high falling tone) or "*pointy*" (level tone). Word formation in Manipuri uses a large number of suffixes with fewer affixes and primarily uses the process of compounding. For example, *lan (war)* + *mee (person)* - "*lanmee*" (*soldier*). Manipuri has two scripts - Bengali and Meitei-Mayek. We are using the Bengali script for our work.

What is more challenging with our work is that Manipuri is a low-resource language. It is known that Neural Machine Translation (NMT) is data-hungry, and Statistical Machine Translation (SMT) is a better option at low resource conditions [2]. The sufficient training data size for NMT is in the order of millions [3] compared to few thousand for SMT. Above this, NLP tools, such as morphological analysers and part-of-speech (POS) taggers, are unavailable for the Manipuri language. This unavailability greatly restricts the researchers from implementing and trying out different possibilities in their area of research. The two techniques, which are pretty conventional in the area of MT, are SMT and NMT - with NMT getting more acclamation than SMT. Research on translations from more affluent to poorer morphology and vice versa is rarely focussed. The study [4] reports that translating from poorer to richer morphology is more complex and challenging than vice versa. However, for translations that involve two languages which are morphologically and structurally variant, determining which technique is better is still a question of doubt. Our work uses the SMT technique to develop a translation system for English-Manipuri. SMT makes use of parallel corpora and learning algorithms to train a model. Based on the translation model's training process, many SMT approach models are available: baseline model, phrase-based model, factored model and hierarchical phrase-based. Traditional or baseline models use word-level mapping, which produces low translations. An improvement to it is the phrasal one, the most routine SMT. The phrase-based SMT (PBSMT) uses a grouping of surface forms of words treated as phrases. On the other hand, a factored model, an extension of PBSMT, uses surface word forms with additional factors such as lemma, POS tags, morphological information, case, and genders. In contrast, the hierarchical model requires grammar which consists of Synchronous Context Free Grammar rules. In SMT, language models are used to establish the target word order. They are, however, limited by the sparsity of the data caused by larger n-grams. Therefore, a lexicalised reordering model subjects reordering to the PBSMT phrases. One of the most challenging issues in SMT is reordering; it manifests differently depending on the language combinations. Language pairings with far-off syntactic structures, like English and Manipuri, experience long-distance reordering issues that the lexicalised reordering models cannot resolve.

In our paper, we implement a factored model for English-Manipuri along with pre-processing, post-processing, and reordering modules. The following sections give a detailed explanation of the architecture. Unfortunately, unlike high-resource languages such as English with POS taggers available in the open domain, no such tool is available for Manipuri. It is also difficult to find annotated corpora for this language. So, owing to financial reasons, we have developed only a small set of annotated corpora of the entertainment domain to experiment with the system. Additional corpora with more factors, if available, can be incorporated to improve the result. This work is a quest to improve translation quality.

The paper implements a factored model with POS information-based reordering to enhance translation between linguistically disparate language pairs. The goal of the current task is to rearrange the source chunks so that the alignments of the source and destination pieces are more monotonous. The reordering has been done by manually rearranging the source text at the chunk level to replicate the ordering of the target language. The process mitigates the problem of long-range reordering to only short-range, intra-chunk reordering that the lexicalised reordering easily handles. Furthermore, Manipuri's low resources and agglutinating nature produces untranslated words, which are transliterated.

2. FACTORED SMT

The most dominant SMT approach, the PBSMT model, has an extension model called the factored SMT model. The phrase model uses small text chunks and phrases without linguistic information during translation. The PBSMT model uses the noisy channel model. After applying Bayes theorem [5], the translation probability for translating a source sentence (S) to a target sentence (T) is

$$P(T/S)=P(S/T)P(T)=argmax_m\, P(S/T)P(T) \qquad (1)$$

In Equation 1, the component P(S/T) represents the translation model. In contrast, $P(T)$ represents the language model. For finding the best translation of a given source sentence, S we use a decoder to find the best probable target sentence, T. The decoder finds the n-best possible translations of S to T, out of which the translation with the highest probability is chosen, specified by $argmax_m$ in Equation 1.

The phrase-based technique provides promising results for language pairs that are structurally and morphologically similar. In the PBSMT framework, the distortion and lexicalised reordering models are widely used to handle reordering. The lexicalised reordering of Moses makes reordering simple for language pairs with analogous syntactic structures. However, the translation is quite bad for language pairs that are structurally and morphologically distant from each other with low resources, due to the problem of long-distance reordering. Although the basic PBSMT model is expandable to account for long-distance reordering, it often performs worse due to distortion limitations [6]. Mapping at the grouping of surface forms of words does not reflect the pattern between the languages. Therefore, it is difficult for the PBSMT model to learn the translation pattern between the languages. To deal with it, the factored SMT model is adopted. In the factored model, a word is represented not only by its surface form, but by multiple levels called factors - such as lemma, POS, and morphological information. We are incorporating linguistic features into the corpus with this multiple-level representation of words. Incorporating linguistic information through factors will aid in learning a translation model that addresses the linguistic divergences between the languages. Works using factored SMT models significantly improve the translation result of PBSMT. One such implementation is for the Kannada language [7]. The factored model of SMT is mainly helpful in involving language pairs where one is morphologically rich and the other poor. We can feed more factors to improve our translation results better. The contributions of our work are:-

- We develop a small set of POS-tagged Manipuri corpora of the entertainment domain using the IL-POST (Indian Language Parts of Speech Tagset) [8] framework.
- We implement a PBSMT model as the basis for comparison.
- We implement factored SMT models with the inbuilt reordering option of Moses.
- We perform hand-coded reordering of the POS-tagged English side sentences of the training and testing corpus and implement a factored model of SMT.
- We compare the results of our systems using automatic evaluation metrics and establish that our architecture improves translation results.

3. LITERATURE REVIEW

Works that integrate linguistic information into the more routine PBSMT are limited. Here, we discuss some works that use the factors and factored SMT model in different ways to improve the translation result and address data sparsity, grammatical error, and fluency for morphologically rich languages in MT.

They describe one of the early approaches for translating French to English using linguistic information in work [9]. Utilising factored models for English-Latvian and English-Lithuanian SMT systems is the subject of yet another implementation report [10]. Latvian and Lithuanian languages are highly inflectional. They belong to morphologically rich, have free phrase order and are highly ambiguous, which results in data sparseness in translation. They have addressed this issue by splitting each token into its stem and suffix parts and treating them as separate models for Lithuanian-English translation. While for English-Latvian, morphologic tags are used as an additional language model apart from suffixes. Their work claims a significant improvement over baseline SMT through human evaluation. Similar experiments with phrase-based MT for English-Czech were conducted by [11], demonstrating the benefits of utilising multiple factors. His work involves different models involving combinations of word forms, lemma and morphological tags as factors. His work concludes with the BLEU (Bilingual Evaluation Understudy) score report demonstrating that multi-factor SMT consistently outperforms baseline SMT. In another study, a factored SMT model is implemented by [12] using fixed-length word suffixes that approximate POS tags in some ways. Their work reduces the language model's perplexity and increases the grammatical correctness of the results. Their work shows an improvement over the baseline SMT.

Contrary to translations from morphologically rich languages, translations to them have the limitation of lousy translation quality output. One major issue with morphologically rich languages is the data sparseness problem. Various reports associated with data sparseness are available for morphologically rich languages such as Latvian, Lithuanian, Croatian, Tamil, Malayalam, Mizo, Hindi, Kannada, and Farsi. Their work handles the data sparsity problem in translating a morphologically rich language [13]. They suggested a solution that generates unseen morphological forms fed into the training corpora. Their proposed solution claims to improve the translation quality through translation experiments of English to Hindi and Marathi languages.

Works that perform pre-processing, apart from having factors on the corpus, are available. One such factored SMT system for English to Tamil [14] is available. Their model uses lemma, POS and compound tag as factors on the source side; and lemma, POS tag and morphological information on the target side. They develop a novel pre-processing approach on the source language (English) so that it conforms to the target language (Tamil). The training uses the pre-processed sentences using a factored SMT model. Finally, morphological generators of the Tamil language generate a surface form of words from the factors output by the SMT model. The output result outperforms Google Translate and other systems. Pre-processing to change the source sentences' (English) structure by adding POS tags is another similar effort [15]. They use POS tags to modify the English sentences to be more similar to the richer Spanish and Catalan target sentences. However, some rely upon post-processing rather than pre-processing. In their work [16], they adopt post-processing techniques using syntactic and morphological knowledge of both source and target data to predict inflected forms of a sequence of word stems of the target side.

The work of [7] compares the factored model and the baseline model of the SMT technique for the morphologically rich Kannada translation. They create language models based on surface form and POS tag for the factored model. They report that the factored model provides an improvement of 25% in BLEU compared to the baseline model. In another paper [17], morphological information factors comprising word stem, prefix, suffix, and POS tag - handle grammatical error correction. For evaluation, they modelled five individual systems with each factor - stem, prefix, suffix, and POS tag. Evaluation of the systems uses measuring their performance in the grammatical error correction task. Their models show an improvement in BLEU by 32.54% over the phrase-based model. Further, their model was experimented with official test data and compared with thirteen other systems at the "CoNLL 2014 shared task", in which they got the 7th and 5th F0.5 scores. They concluded that POS information is the most effective, out of all factors, and that the model with POS information outperforms others with other factors. Even with the latest technique of NMT, [18] they apply pre-ordering to reduce word order divergence between source and target languages for a few resourced Indic languages. Their approach relatively improves the translation quality even in low-resource scenarios.

For the Manipuri language, [19] they work reports the only factored SMT model implementation. They have used suffix and dependency relations as factors on the source (English) side and case markers on the target (Manipuri) side. Their system was trained on 10,350 sentences and tested on 500 news domain datasets. BLEU score and subjective evaluation claim an improvement in their result. MT systems for the Manipuri language are getting built but are far from perfect. Currently there isn't any Manipuri-language content that makes use of POS tags or sentence restructuring. Their study [20] evaluates the effectiveness of unsupervised MT models for Manipuri-English translation using a comparable corpus of news domain. They use a suffix segmenter using graph-based stemmer and transliteration models to re-score the sentence translation and the lexical probability. They report that the unsupervised SMT model is more successful than the unsupervised NMT models for the language pair. Using data augmentation approaches, they addressed low-resource problems while experimenting with a semi-supervised approach [21]. The data augmentation process uses comparable monolingual corpora from the news domain. They employ a self-training and back-translation approach to produce synthetic parallel data from monolingual data. According to reports, their model outperforms unsupervised, supervised, mBART, and other standard semi-supervised models in quantitative efficacy and can handle sparse data. Additionally, they make empirical claims about how well their models cope with uncommon words and long-term dependencies.

Their study [22] uses the multilingual pre-trained models mBART50 and mT5-base with fine-tuning for transfer learning of low-resource MT involving Assamese, Manipuri, and Bengali languages. Their fine-tuned models outperform the multilingual baseline model indicated by their BLEU scores. On the WAT-2021 test set, their model with mT5-base fine-tuning performs best. Their model, however, predicts a lower score for the Flores-101 test set. Another experiment with a many-to-many NMT model with cross-lingual capabilities is reported [23]. Their work enhances the basic paradigm for many-to-many translation between Manipuri and English and the bilingual model. Additionally, they use zero-shot translation to examine the generalizability of their methodology on language pairs with no direct correspondence and contrast it with pivot-based translation. To translate English to Manipuri, [24] examine the supervised and unsupervised SMT and NMT techniques. They also test out low-resource techniques like self-training and back translation. They examine the difficulties and mistakes made in translating English to Manipuri. Works of [25], [26] and [27] are a few other NMT implementations for the Manipuri language. Recently, Google Translate [28] supported Manipuri using the Meitei-Mayek script, using the concept of Zero-Shot translation [29].

4. PROBLEM STATEMENT

SMT and NMT are the techniques that dominate MT. Dealing with dissimilar language pairs in low-resource settings is still challenging for both techniques. Based on the language pair under consideration, some systems favour SMT [30], while some favour NMT [31]. SMT is good at handling adequacy but trades for fluency [2], while NMT trades adequacy for fluency [32]. NMT has replaced most larger MT systems, which have larger corpora. However, SMT has suggested a better option for low-resource conditions. SMT uses mathematical models to map the source and target language symbols. The mapping is more straightforward for similar languages, and the results are promising. Despite the state-of-the-art PBSMT's usage of reordering models, the inconsistency in word ordering between distant languages leads to subpar translation quality. We pose the divergences between languages and the inability of lexicalised reordering to handle long-distance reordering as a problem. We propose prior source-side reordering to mitigate the syntactic differences and handle long-distance reordering to improve the translation result.

5. PROPOSED METHODOLOGY

As mentioned above, English and Manipuri languages have significant linguistic differences in comparison. Their morphological and structural differences increase the challenges of translation. To treat this, we first employ separate pre-processing modules for both English and Manipuri sentences, followed by training. System training uses our proposed model, followed by a transliteration module, for handling untranslated tokens present in the output. Further explanation is in the following subsections.

5.1. PRE-PROCESSING

5.1.1. English Sentences

The pre-processing of English sentences follows the order of tokenising, factoring and reordering. Tokenising of the English side uses the inbuilt tokeniser of Moses. A factoring module then treats the tokenised output to extract the factors. The factors may be lemma, POS and morphology. Here, we use the maximum entropy-based MXPOST [33] framework to include only POS tags as factors, as the target side has only POS tags as a factor. The MXPOST uses the Penn treebank POS tagset. Using the POS information, we reordered the factored sentences using the linguistic rules of Manipuri. Reordering transforms the structure of English sentences to make them structurally more similar to Manipuri. This transformation will reduce the syntactic divergences between the language pair, thereby better mapping between source and target symbols. The source text's prior reordering has two effects on how well the MT performs [34]. Firstly, it handles long-distance reordering and thus, reduces the workload on the reordering model by prior reordering the source text. Only trivial reordering occurs during decoding, and the translation hypothesis's construction uses a monotone orientation. Secondly, prior source reordering should result in more accurate word alignments, better translation models and higher translation quality since statistical word alignment approaches function efficiently for linguistic groups with analogous grammatical structures. Fig 1. represents the pre-processing module of English sentences. As an example, we have an English sentence below.

Example:

(Before pre-processing)

I am a boy.

(After factoring)

I|PRP am|VBP a|DT boy|NN .|.

(After reordering)

I| PRP boy| NN am| VBP a|DT .|.

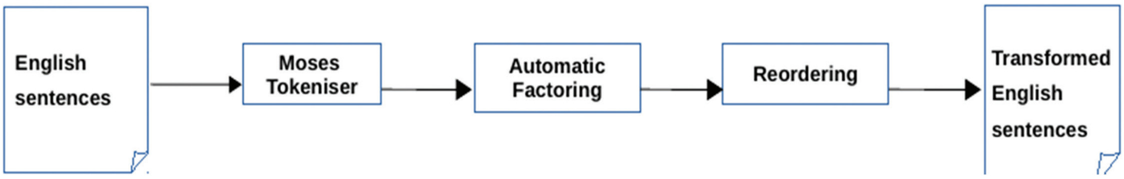

Fig. 1. English side pre-processing

5.1.2. Manipuri Sentences

Fig. 2 shows the pre-processing module of Manipuri sentences. The process is similar to English for Manipuri side pre-processing, except the reordering step is not present. The tokenising step here uses the dedicated indicNLP [29] tokeniser for the Manipuri language.

Tokenising is followed by factoring. Higher the number of factors, the more refined our system is. For our current work, we have considered only the POS tag as the factor due to the high cost of manual preparation, as open-source NLP tools are not available for Manipuri. We manually develop our annotated corpora of the entertainment domain using the ILPOST for the Manipuri language.

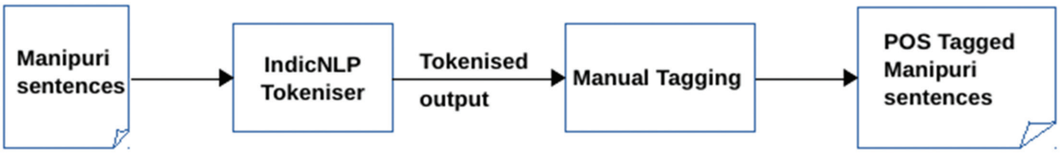

Fig. 2. Manipuri side pre-processing

5.2 Architecture of the System

The overall architecture of our system is in Fig. 3. We use the outputs of pre-processing modules of Fig. 1 and Fig. 2 as the dataset to train our system. There are no POS taggers or annotated corpora for the Manipuri language. Consequently, we developed the dataset for this research with the aid of a linguist. It is essential to specify that the dataset we use to train our system is relatively small owing to financial reasons. We split up the dataset into training, development, and test set using Python's split code. We employ a training set of 8,000 sentences, a development set of 1,000 sentences, and a test set of 2,000 sentences from the entertainment domain. The source side test data undergoes the pre-processing of Fig.1 before feeding for translation. Therefore, the translation step in our system uses the pre-processed text. Adequacy measures the mapping between source and target symbols, which is not an issue for translations involving language pairs with the same structure and morphology. However, when the structure and morphology are different, it is a problem that needs addressing. The language divergence makes it more challenging to map source and target symbols correctly, affecting the adequacy and fluency of translation results, which are the measures of translation quality.

Our architecture uses factoring followed by hand-coded reordering on the source side to address this issue. Through reordering, we attempt to reduce the structural divergences; thus, the translation output will closely resemble the target side language, thereby providing a better mapping mechanism between source and target symbols. Furthermore, when using a factored model, we consider the POS feature in addition to the surface form to produce the translated output. Therefore, incorporating a POS tag on the dataset provides language-specific linguistic knowledge to the training model. Lastly, given the small training data, untranslated words will appear in the translated output, which uses transliteration as a post-processing step.

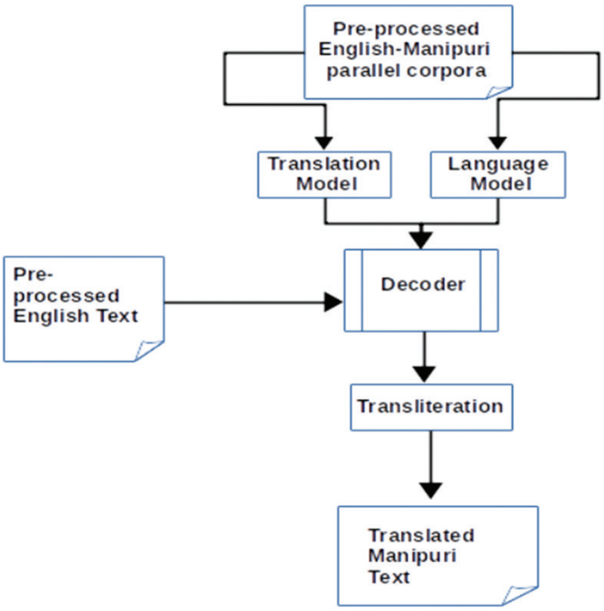

Fig. 3. Architecture of our proposed model of factored English-Manipuri system

6. RESULTS AND DISCUSSION

Our language pair, as mentioned above, has different word order. The distortion limit limits the distance between the following and previously translated phrases, which might seriously impair disparate languages. We, therefore, set the distortion limit to 0.5 instead of the default 0.3 to increase the search space of the decoder. For comparison, we experiment with multiple systems with different parameter settings using the same data size. One system is the general PBSMT which we use as a basis. This system data does not have any factoring or explicit reordering involved.

We also train three different factored models. FSMT1 (factored model with word-based reordering), FSMT2 (factored model with phrase-based reordering) and FSMT3* (factored model with hand-coded reordering on source side). The FSMT3* is the model which uses our proposed architecture model of Fig. 3. It is essential to mention that in FSMT3*, before translation, the source side of the test set first undergoes factoring, followed by hand-coded reordering.

Table 1. Score comparison of the systems

	PBSMT	FSMT1	FSMT2	FSMT3*
BLEU	3.58	3.47	3.71	4.12
F1	0.181	0.181	0.183	0.193
Meteor	0.098	0.093	0.10	0.1
Precision	0.170	0.175	0.171	0.187
Recall	0.194	0.189	0.195	0.2

Table 1 shows the BLEU [35], F1, Meteor [36], Precision and Recall scores of these systems tested using a test set size of 2000 sentences. We find that PBSMT outperforms FSMT1 in terms of BLEU, Meteor, and Recall scores, except Precision - which suggests that phrasal reordering obscures FSMT1's POS feature information. Therefore, the FSMT2 model, which utilises both POS characteristics and lexicalised phrasal reordering, performs better in all scores than PBSMT and FSMT1. However, the proposed model, FSMT3*, proves a further improvement in all scores, even over the FSMT2 model discussed above. FSMT3* outperforms the FSMT2 model by 11.05% (BLEU), 5.46% (F1), 9.35% (Precision), and 2.56% (Recall), even with scant training data. Our result shows that reordering the English sentences as per the Manipuri syntax along with POS features improves the translation quality, even when the dataset is small.

7. CONCLUSION

Our work found that handling linguistic divergences is lucrative in MT. Pre-ordering the source side and adding POS as a linguistic characteristic increased the scores by 11.05% (BLEU), 5.46% (F1), 9.35% (Precision), and 2.56% (Recall), respectively, from that of FSMT2. Despite being small, this gain represents a significant improvement given the dataset and feature limitations. Prior reordering handles long-distance reordering, thus mitigating language divergences. Furthermore, language-specific hand-coded reordering of the source side chunks that match the target language chunks provides better alignment than the lexicalised reordering option of Moses. It thus improves translation quality even under low-resource settings.

Our current work uses a small dataset prepared for experimental purposes, and reordering is also manual. If Manipuri-specific NLP tools are available, we can automate the factoring and reordering process by incorporating other features to refine our results further. Our hand-coded reordering model on the source side with POS as a feature for the low-resource English-Manipuri translation is a bootstrapping strategy towards reducing the linguistic gap for enhancing translation quality. The goal is to find techniques to close the linguistic divergence gap because MT models never produce the best word alignments for languages with far-off linguistic features. Due to our study's usage of different scripts, we exclude a comparison with the Google Translate result.

8. REFERENCES

[1] P. Passban, "Machine Translation of Morphologically Rich Languages Using Deep Neural Networks", PhD diss, Dublin City University, 2017.

[2] S. Sen, M. Hasanuzzaman, A. Ekbal, P. Bhattacharyya, A. Way, "Neural machine translation of low-resource languages using SMT phrase pair injection", Nat. Lang. Eng., Vol. 27, No. 3, 2021, pp. 271–292.

[3] G. Lample, M. Ott, A. Conneau, L. Denoyer, M. Ranzato, "Phrase-Based & Neural Unsupervised Machine Translation", arXivarXiv preprint arXiv:1804.07755, 2018.

[4] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation", Proceedings of Machine Translation Summit X: Papers, Phuket, Thailand, September 2005, pp. 79–86.

[5] J. V. Stone, "Bayes' Rule: A Tutorial Introduction to Bayesian Analysis", Sebtel Press, 2013.

[6] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, "Moses: Open Source Toolkit for Statistical Machine Translation", Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, Prague, Czech Republic, June 2007, pp. 177–180.

[7] K. M. Shivakumar, N. Shivaraju, V. Sreekanta, D. Gupta, "Comparative study of factored SMT with baseline SMT for English to Kannada", 2016 International Conference on Inventive Computation Technologies (ICICT), August 2016, vol. 1, pp. 1–6.

[8] LDC-IL, https://www.ldcil.org/standardsTextPOS.aspx (accessed: 2022)

[9] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, J. D. Lafferty, R. L. Mercer, "Analysis, statistical transfer, and synthesis in machine translation", Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Montréal, Canada, June 1992.

[10] A. Utka, J. Vaičenonienė, J. Kovalevskaitė, "Human Language Technologies – The Baltic Perspective: Proceedings of the Ninth International Conference Baltic HLT 2020", IOS Press, 2020.

[11] O. Bojar, "English-to-Czech Factored Machine Translation", Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic, June 2007, pp. 232–239.

[12] N. Sharif Razavian, S. Vogel, "Fixed Length Word Suffix for Factored Statistical Machine Translation", Proceedings of the ACL 2010 Conference Short Papers, Uppsala, Sweden, July 2010, pp. 147–150.

[13] P. D. Dungarwal, "Reordering Models for Statistical Machine Translation: A Literature Survey", Indian Institute of Technology, Bombay, India, 2014.

[14] A. Kumar, "Factored Statistical Machine Translation System for English to Tamil Language", Pertanika Journal of Social Sciences & Humanities, Vol. 22, No. 4, 2014.

[15] N. Ueffing, H. Ney, "Using POS Information for SMT into Morphologically Rich Languages", 10th Conference of the European Chapter of the Association for Computational Linguistics, Budapest, Hungary, April 2003.

[16] E. Minkov, K. Toutanova, H. Suzuki, "Generating Complex Morphology for Machine Translation", Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, June 2007, pp. 128–135.

[17] R. Wang, C. Ding, M. Utiyama, E. Sumita, "English-Myanmar NMT and SMT with Pre-ordering: NICT's Machine Translation Systems at WAT-2018", Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation, 2018.

[18] R. Murthy, A. Kunchukuttan, P. Bhattacharyya, "Addressing word-order Divergence in Multilingual Neural Machine Translation for extremely Low Resource Languages", Proceedings of the 2019 Conference of the North, Minneapolis, Minnesota, 2019, pp. 3868–3873.

[19] T. D. Singh, S. Bandyopadhyay, "Manipuri-English Bidirectional Statistical Machine Translation Systems using Morphology and Dependency Relations", Proceedings of the 4th Workshop on Syntax and Structure in Statistical Translation, August 2010, pp. 83-91.

[20] L. Laitonjam, S. Ranbir Singh, "Manipuri-English Machine Translation using Comparable Corpus", Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021), Virtual, August 2021, pp. 78–88.

[21] S. M. Singh, T. D. Singh, "Low resource machine translation of english–manipuri: A semi-supervised approach", Expert Systems with Applications, Vol. 209, 2022, pp. 118-187.

[22] S. M. Singh, L. Sanayai Meetei, A. Singh, T. D. Singh, S. Bandyopadhyay, "On the Transferability of Massively Multilingual Pretrained Models in the Pretext of the Indo-Aryan and Tibeto-Burman Languages", Proceedings of the 18th International Conference on Natural Language Processing (ICON), National Institute of Technology Silchar, Silchar, India, December 2021, pp. 64–74.

[23] S. M. Singh, T. D. Singh, "An empirical study of low-resource neural machine translation of manipuri in multilingual settings", Neural Computing and Applications, Vol. 34, No. 17, 2022, pp. 14823–14844.

[24] T. J. Singh, S. R. Singh, P. Sarmah, "English-Manipuri Machine Translation: An empirical study of different Supervised and Unsupervised Methods", 2021 International Conference on Asian Language Processing (IALP), December 2021, pp. 142–147.

[25] S. M. Singh, T. D. Singh, "Statistical and Neural Machine Translation Systems of English to Manipuri: A Preliminary Study", Soft Computing and Signal Processing, Singapore, 2021, pp. 203–211.

[26] L. Rahul, L. Meetei, H. Jayanna, "Statistical and Neural Machine Translation for Manipuri-English on Intelligence Domain", Advances in Computing and Network Communications, pp. 249–257, Springer Singapore, 2021.

[27] S. M. Singh, T. D. Singh, "Unsupervised Neural Machine Translation for English and Manipuri," Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages, December 2020, pp. 69-78.

[28] Google Translate, https://translate.google.co.in (accessed: 2022)

[29] Google Translate adds support for Assamese, Mizo and Manipuri languages - Eastern Mirror, https://easternmirrornagaland.com/google-translate-adds-support-for-assamese-mizo-and-manipuri-languages (accessed: 2022)

[30] M. Dowling, T. Lynn, A. Poncelas, "SMT versus NMT: Preliminary comparisons for Irish", Association for Machine Translation in the Americas (AMTA), 2018.

[31] S. Kinoshita, T. Oshio, T. Mitsuhashi, "Comparison of SMT and NMT trained with large Patent Corpora: Japio at WAT2017", Proceedings of the 4th Workshop on Asian Translation (WAT2017), November 2017, pp. 140–145.

[32] P. Koehn, R. Knowles, "Six Challenges for Neural Machine Translation", arXiv preprint arXiv:1706.03872, 2017.

[33] A. Neumann, nltk-maxent-pos-tagger, https://github.com/arne-cl/nltk-maxent-pos-tagger/blob/023241f9deceeb214cef7304b5a8ebc914024dfd/mxpost.py (accessed: 2022)

[34] M. Holmqvist, S. Stymne, L. Ahrenberg, M. Merkel, "Alignment-based reordering for SMT", Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, May 2012, pp. 3436–3440.

[35] Tilde MT, https://www.letsmt.eu/Bleu.aspx (accessed: 2022)

[36] The METEOR Automatic MT Evaluation Metric, http://www.cs.cmu.edu/~alavie/METEOR/ (accessed: 2022)