AN ANALYSIS OF PHRASE BASED SMT FOR ENGLISH TO MANIPURI LANGUAGE

Maibam Indika Devi[1] and Bipul Syam Purkayastha[2]

[1]Department of Computer Science, IGNTU-RCM, Kangpokpi, Manipur, India
[2]Department of Computer Science, Assam University, Silchar, Assam, India

ABSTRACT

Statistical Machine Translation (SMT) is one ruling approach adopted for developing major translation systems today. Here, we report a phrase-based SMT system from English to Manipuri. The variance in the structure and morphology between English and Manipuri languages and the lack of resources for Manipuri languages pose a significant challenge in developing an MT system for the language pair. In comparison, English has poor morphology and SVO structure and belongs to the Indo-European family. Manipuri language has richer morphology and SOV structure and belongs to the Sino-Tibetan family. Manipuri has two scripts- Bengali script and Meitei script. Here the Bengali script is used for developing the system. Our system uses the Moses toolkit. We train and test the system using the tourism, agriculture and entertainment corpus. Further, we use the BLEU metric to evaluate the systems' performance.

KEYWORDS

Phrase-based SMT, English- Manipuri, Moses, BLEU.

1. INTRODUCTION

Machine Translation (MT) is an important area in Natural Language Processing (NLP) where many systems are being developed worldwide for translation from one language to another. It aids in the translation process, be it a book, movies, official documents from one language to another. English is a simple and easy to learn language. Most documents, books, journals, articles, web pages are available in English. However, with the application of MT, articles or web pages can be viewed in a different language. There are various techniques to MT of which Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) is the most prominent. SMT is the technique where translation is done through statistical models.. In a phrase based SMT model, the translation units are phrases rather than words. To perform translation, phrases in source language will be mapped with target language phrases, by using maximum likelihood estimate; the best translation out of many candidate translations will be selected. We use the open source toolkit Moses[1] to implement the phrase based model of SMT technique.

English is the source language, and Manipuri is the target language. The language pair treated here, English and Manipuri, is a challenging one because of the huge difference in terms of linguistic structure. Manipuri belongs to the Sino-Tibetan family, has SOV structure, tonal, rich morphology, aspect predominance and synthetic category and agglutinating. While English comes under the Indo-European family, has SVO structure, stressed and non-tonal, poor morphology, tense dominant and analytic category. The huge difference in the language structure

David C. Wyld et al. (Eds): CST, NLMLT, DMS, CLBD, ITCS, VLSIE - 2022
pp. 37-43, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121904

provides a challenge while performing translation. In addition to this, there is a lack of reliable pre-processing tools and resources for the Manipuri language. Though some works are seen for Manipuri language, the existing tools are not satisfactory and reliable.

There are also very limited resources for the Manipuri language. Only a few sentences of bilingual data are available to the public for research. These data are not enough for efficient working and quality output of the system. So, limited corpus and tools, difference in structure and morphology of language under consideration pose challenges in developing the translation system.

The bilingual and monolingual corpora used for training the system consist of varying domains from tourism, agriculture and entertainment. Some of these corpora are downloaded from TDIL[2] website while some are manually developed. For each domain, we train and test separate phrase based SMT systems. We evaluate how the system performs for these domains having different data sizes using automatic evaluation metric.

2. RELATED WORK

[3] has written a survey paper on the various approaches to machine translation and the major translation systems developed for Indian languages. Most of the major Indian languages has well developed machine translation systems. The general purpose Google Translate[4] provides good results. However when dealing with specific domain related translations, tailor made MT systems trained on that domain will better serve its use.

The North-East section of India has a diversity of languages with multiple dialects. [5]has discussed the works carried out in NLP for north-eastern languages covering Hindi, Manipuri, Assamese, Kokborok, Nepali, Mizo, Bodo, Bengali etc. Nevertheless, NLP related advancements are found in the works of Assamese[6], Nepali [7], Bodo[8]. An open-access NLP toolkit[9] dedicated to Bengali is available also. In comparison to them, Manipuri language is lagging far behind. Even though recognized by the Indian Union as one of the scheduled languages, there is little work in NLP applications. The non-availability of resources, language characteristics, and lack of experts poses some of the factors that hinder its development. The survey report[10] covers areas on E-dictionary, Machine Translation, POS tagging, WordNet, Word Sense Disambiguation, Multi-word expressions, Name Entity Recognition, Morphological Analysis. Some of the Manipuri language related MT works are as follows.

[11]developed the Manipuri-English Example-based Machine Translation. The corpora used here is of news domain with POS tagging, NER, morphological analysis and chunking applied. They have measured the output using BLEU and NIST metrics, scoring 0.317 and 3.361, respectively, depending on which claims has been made that the EBMT approach is better than baseline SMT on using the same set of data. [12] developed Manipuri-English Bidirectional SMT systems. Their system used a corpus from the news domain with 10350 sentence pairs for training and 500 sentences for testing. Apart from using the statistics of the corpus, they have incorporated additional morphological information into the system. The English-Manipuri pair has incorporated suffix dependency relations on the source side and case markers on the target side. While for Manipuri-English pair, case markers, POS tags on the source side and suffix and dependency relationships on the target side. Both the translations showed improved results from the baseline system as given by their BLEU score. [13] developed the factored SMT for the English-Manipuri language pair. Suffix and dependency relations are treated as factors on the source side and case markers on the target side. The system is trained using 10350 sentences and tested on 500 sentences. The output shows an improved BLEU score. [14] has carried out the

Phrase-based SMT for Manipuri languages by integrating reduplicated MWE. They have stated that the integration improves the BLEU and NIST scores over baseline SMT. [15] has carried out machine translation from English to Manipuri using SMT and NMT. The output comparison shows NMT having a higher BLEU score as compared to phrasal SMT. In their work "Unsupervised Neural Machine Translation for English and Manipuri" they reported a BLEU score of 3.1 for English-Manipuri translation and 2.7 for Manipuri-English translation. [16] developed the Manipuri-English translation system using the intelligence domain. They used a corpora of 56,678 size from the intelligence domain based on open-source intelligence (OSINT). Evaluation of SMT and NMT is done based on BLEU score, where NMT outperforms SMT. They also incorporated suffix based morphological analysis information which further improves the BLEU score.

3. DEVELOPING THE SYSTEM

3.1. Corpus Preparation

Parallel corpora are a collection of sentences of two different languages which are aligned at the sentence level. The bilingual parallel corpus will be used to train the system. It is the quality and quantity parallel corpus fed into the system that characterizes the result of translations. Therefore bigger the corpora better the system performance. Tourism corpus from TDIL[2] along with newly developed corpus of entertainment, tourism and agriculture are used for training. Table 1 shows the corpus distribution of the different domains used in developing the system.

Table 1. Corpus distribution of different domains used in training the system.

Domain	Translation Model	Language Model
Agriculture	10,000	500
Entertainment	10,000	500
Tourism	25,000	1000

3.2. Preprocessing

The preprocessing steps include tokenizing, true casing and cleaning of parallel data.Tokenizing is the step for identifying tokens such as words, numbers, and punctuations. We use the inbuilt tokenizer for English sentences. Moses inbuilt tokenizers have no support for Manipuri language. So, we use the IndicNLP tokenizer[17]. After this, we perform truecasing and cleaning of the tokenized output. In the cleaning step, we set the length limit to 80.

3.3. English To Manipuri System

In SMT, a source language sequence 'e' (English) is translated into a target language sequence 'm', by computing the most likely translation using the following equation[18] ,

$$p(e,m) = argmax_m p(m|e)(Equation1)$$

Using Bayes Rule [19], Equation 1 is written as-

$$argmax_m p(m|e) = argmax_m p(e|m)p(m)(Equation2)$$

In Equation 2, the component p(m|e) in Equation 1 is decomposed into two components. The component p(m) is the language model, and another component p(e|m) is the translation model,

which is discussed in the latter part of the paper. The English to Manipuri MT system is developed using the phrase-based SMT technique. In phrase-based SMT, the translation units are phrases. A foreign English sentence is segmented into phrases, and each English phrase is mapped into Manipuri phrases. The phrases can also be reordered. As compared to baseline SMT where the translation units are words, the phrase-based SMT provides better results. Various toolkits are available to implement the SMT model of machine translation, Moses being one of them. Moses is open sourced and the most commonly used toolkit for developing SMT systems. The two main components of Moses, the training pipeline and the decoder, form the basis for translation. Apart from this, Moses consists of multiple tools and utilities and also supports various external tools. Developing a translation system from training data requires multiple stages, where the stages are implemented in a pipelined manner, hence the name training pipeline. Moses provides the advantage to add various external tools during the training pipeline. However, the parallel corpora is not used directly for training the system. They are preprocessed first.

3.4. Language Model

In Equation 2, the component p(m) represents the language model. Only the monolingual target side corpus (Manipuri corpus) is required to create the language model. The size of the monolingual corpus used here is separate from that used in training. The language model makes the translation system aware of how the target language should appear and ensures fluent output. For creating language models, the following monolingual corpora are used.

1. Monolingual Manipuri sample general corpus from TDIL.
2. Monolingual Manipuri sample raw corpus from NPLT [20].

Moses supports various language modeling tools such as KenLM, IRSTLM, SRILM and RandLM. Here, the built-in KenLM model is being used. Here, a 3-gram modeling technique is used to compute the probability of Manipuri sentences, denoted by p(m). The component p(m) is calculated based on Markov's Chain Rule [18] as,

$$p(m) = \Pi_{i=0}{}^{m} \quad p(m_i|m_{i-1}, m_{i-2})(Equation3)$$

Where m_i is the current word generated.

3.5. Translation Model

The component p(e|m) in Equation 2 is the translation model. As the component shows, bilingual parallel corpora are involved in creating the translation model. It estimates the lexical correspondence between the languages. The translation model computes the probability of a source sentence for a given target sentence and tries to find the best translation of a given phrase. The probability p(e|m) is computed as the summation of all probabilities with possible alignment 'a' between the phrases of English and Manipuri language,

$$p(e|m) = \sum_a p(e, a|m)(Equation4)$$

Here, the built-in tool GIZA++ aligns the words between the source and target languages. In phrase based SMT, the translation units are phrases; therefore, the translation model is built based on the frequency of occurrences of phrases in the training corpus. This information is stored in a table called a phrase-table which contains the phrases and their frequency over the entire training corpus the higher the frequency of a phrase, the greater the chances of getting a correct

translation. The phrase table forms the translation model for the system. The role of the translation model is to ensure that the source language and target language are good translations of one another.

3.6. Decoder

The decoder takes input sentences in the source language (English) and uses the translation model and language model to translate them into the target language (Manipuri). The decoder is responsible for determining the best translation out of its many candidate translations. It uses the argmax() function to find the maximum translation probability of all candidate translations. There are many tools for decoding in SMT systems. Here, the inbuilt Moses decoder is used.

4. EVALUATING THE SYSTEM

4.1. BLEU

BLEU (Bilingual Evaluation UnderStudy) is the dominant and language-independent metric for measuring translation quality. [21]BLEU score counts the number of matches in a weighted fashion, the consecutive phrases between the machine-translated output and the reference translations made by humans. Out of different BLEU available, we use multibleu.perl to determine the score.

Table 2. BLEU score of the different systems

Domains	Training size	Test data size	BLEU
Agriculture	10,000	1000	7.03
Entertainment	10,000	1000	6.52
Tourism	25,000	1000	14.59

We keep the size of test data the same for all three systems. The training data size is however different. Running a multi-bleu script gives the score in Table 2. As we can see, the size of training data affects the BLEU score. It is common perception that higher the value of BLEU, the better. Moreover, BLEU is directly dependent on the domain size and test data used for training and testing. In our work, however, we do not compare our BLEU scores with those of other MT reports on Manipuri language. This is due to [22] which stated BLEU scores in between papers cannot be compared directly. And that BLEU scores vary with the MT system change, corpus domain, test data, and language pair. Therefore, it is better not to compare the quality of a system, solely based on its BLEU score.

5. CONCLUSION

In our work, we rate the systems of different domains based on BLEU score only to see their result with the little amount of data at hand. The data set used here is insignificant for a well functioning system, however this paper provides an insight on the practicality of the SMT technique on English-Manipuri translations. In our analysis, we get to see that even for such a small amount of training data, the system provides a good output.

SMT is one of the dominating approaches to MT. Nevertheless, there is a recent shift from SMT to NMT in the paradigm of MT. And for both techniques corpus serves as the backbone for functioning. For the language pair English-Manipuri which has got distant linguistic features, it cannot be assumed unless experiment and compare their results which technique will be more

practical. This paper forms a preliminary basis, towards understanding the feasibility and potential of the phrase based SMT system. Our next work will compare NMT and SMT over the same amount of training and testing data.

ACKNOWLEDGEMENTS

We are grateful to Technology Development for Indian Languages (TDIL), National Platform for Language Technology (NPLT) for providing us the corpus and to Professor Chungkham Yashawanta Singh, Fellow at the Indian Institute of Advanced Studies (IIAS) for his support in developing corpus.

REFERENCES

[1] P. Koehn *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*, Prague, Czech Republic, Jun. 2007, pp. 177–180. Accessed: Sep. 06, 2021. [Online]. Available: https://aclanthology.org/P07-2045

[2] "English-Manipuri Sentences of Tourism Domain." https://tdil-dc.in/index.php?option=com_download&task=showresourceDetails&toolid=456&lang=en (accessed Aug. 17, 2022).

[3] P. Antony, "Machine Translation Approaches and Survey for Indian Languages," *Int J Comput Linguist. Chin Lang Process*, 2013.

[4] "Google Translate." https://translate.google.co.in/ (accessed Aug. 16, 2022).

[5] S. Islam, M. I. Devi, and B. S. Purkayastha, "A Study on Various Applications of NLP Developed for North-East Languages," vol. 9, p. 12, 2017.

[6] R. R. Deka, S. Kalita, M. P. Bhuyan, and S. K. Sarma, "A Study of Various Natural Language Processing Works for Assamese Language," in *Intelligent Techniques and Applications in Science and Technology*, Cham, 2020, pp. 128–136. doi: 10.1007/978-3-030-42363-6_15.

[7] T. B. Shahi and C. Sitaula, "Natural language processing for Nepali text: a review," *Artif. Intell. Rev.*, vol. 55, no. 4, pp. 3401–3429, Apr. 2022, doi: 10.1007/s10462-021-10093-1.

[8] M. Narzary, G. Muchahary, M. Brahma, S. Narzary, P. K. Singh, and A. Senapati, "Bodo Resources for NLP - An Overview of Existing Primary Resources for Bodo," *AIJR Proc.*, Jul. 2021, Accessed: Aug. 17, 2022. [Online]. Available: https://books.aijr.org/index.php/press/catalog/book/115/chapter/1126

[9] S. Sarker, "BNLP: Natural language processing toolkit for Bengali language." arXiv, Dec. 01, 2021. doi: 10.48550/arXiv.2102.00405.

[10] M. I. Devi and B. S. Purkayastha, "Advancements on NLP Applications for Manipuri Language," 2018. doi: 10.5121/ijnlc.2018.7505.

[11] T. D. Singh and S. Bandyopadhyay, "Manipuri-English Example Based Machine Translation System," 2011. /paper/Manipuri-English-Example-Based-Machine-Translation-Singh-Bandyopadhyay/080aa68650d13a770bb7a228c10f17ce31baea21 (accessed Mar. 20, 2021).

[12] T. D. Singh and S. Bandyopadhyay, "Manipuri-English Bidirectional Statistical Machine Translation Systems using Morphology and Dependency Relations," *undefined*, 2010, Accessed: Sep. 06, 2021. [Online]. Available: https://www.semanticscholar.org/paper/Manipuri-English-Bidirectional-Statistical-Machine-Singh-Bandyopadhyay/68d64336fb2ac7d302d3ae45051127484754b174

[13] T. D. Singh and S. Bandyopadhyay, "Statistical Machine Translation of English-Manipuri using Morpho-syntactic and Semantic Information," 2010.

[14] T. D. Singh and S. Bandyopadhyay, "Integration of Reduplicated Multiword Expressions and Named Entities in a Phrase Based Statistical Machine Translation System," 2011.

[15] S. M. Singh and T. D. Singh, "Unsupervised Neural Machine Translation for English and Manipuri," p. 10.

[16] L. Rahul, L. Meetei, and H. Jayanna, "Statistical and Neural Machine Translation for Manipuri-English on Intelligence Domain," 2021, pp. 249–257. doi: 10.1007/978-981-33-6987-0_21.

[17] A. Kunchukuttan, "Indic NLP Resources." Jun. 13, 2022. Accessed: Jun. 25, 2022. [Online]. Available: https://github.com/anoopkunchukuttan/indic_nlp_resources

[18] P. Koehn, F. J. Och, and D. Marcu, "Statistical Phrase-Based Translation," p. 8.
[19] J. V. Stone, *Bayes' Rule: A Tutorial Introduction to Bayesian Analysis*. Sebtel Press, 2013.
[20] "A Gold Standard Manipuri Raw Text Corpus." https://nplt.in/demo/index.php?route=product/product&product_id=1987&search=manipuri (accessed Aug. 18, 2022).
[21] laujan, "Legacy: What is a BLEU score? - Custom Translator - Azure Cognitive Services." https://docs.microsoft.com/en-us/azure/cognitive-services/translator/custom-translator/what-is-bleu-score (accessed Aug. 30, 2022).
[22] M. Post, "A Call for Clarity in Reporting BLEU Scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, Belgium, Brussels, 2018, pp. 186–191. doi: 10.18653/v1/W18-6319.