Phrase-Based English–Nyishi Machine Translation

Nabam Kakum and Koj Sambyo

Abstract Machine translation (MT) is the sub-domain of natural language processing (NLP). It process and analyze natural language data which eliminates the language inconceivable matters and help to interact among people of different linguistic backgrounds. In our work, we used the statistical machine translation (SMT) approaches to comprehensively explore the translation accuracy, fluency, and adequacy with the context of low resources Indian dialect (language) Nyishi. SMT needs less training time and works well with highly complex long sentences than other methods, but it requires adequate parallel corpus, which is troublesome in the context of low-resource language like Nyishi. In this paper, we train 30,000 newly collected pairs of corpora and measured the translation accuracy in both directions forward and backward. Finally, results of individual n-gram of BLEU and NIST with and without tuning are calculated, and by using these results, we find out the effectiveness of translation accuracy in respect of fluency and adequacy.

Keywords SMT · BLEU score · Human evaluation NIST · Moses

1 Introduction

MT with its automatic translation method is highly helpful in a multilingual country like India which has 22 scheduled languages. Several researchers tend to work on MT using different methods to eliminate the language barrier issues with the functionality of NLP language translation method and analyze the translated text to obtain better translation accuracy. This work is an experiment to contribute an intelligible translation for the low-source Indian language which is used by the Nyishi people of Arunachal Pradesh. Arunachal Pradesh is inhabited by one of the most culturally and linguistically diverse communities. The ethnic composition of the tribes predominantly belongs to the mongoloid stock. Arunachal Pradesh is the home of 26 major tribes, and more than 100 sub-tribes, Nyishi is considered to be the largest tribe of the

N. Kakum · K. Sambyo (✉)
Department of Computer Science and Engineering, NIT Arunachal Pradesh, Jote, India
e-mail: nabam.phd19@nitap.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

D. Gupta et al. (eds.), *Pattern Recognition and Data Analysis with Applications*,
Lecture Notes in Electrical Engineering 888,
https://doi.org/10.1007/978-981-19-1520-8_38

state with a population of 300,000 approximately from the total population of 13.82 lakhs. Generally, the languages in Arunachal Pradesh belong to the Sino-Tibetan language family and more specifically under the Tibeto-Burman group of languages. This paper attempts to expose low-resource Nyishi Language into machine translation environment by using the properties of NLP with SMT. Many highly resourced languages are already benefitted in almost perfect translation in MT, but the language of the low resource like Nyishi might take many years to advance into MT communities. Nyishi doesn't have a written script which is the main cause of late development and evolvement into MT communities. According to the Worlds languages in Danger (2009) by UNESCO atlas, more than 26 languages of Arunachal Pradesh have been identified as endangered languages as well Nyishi language comes under definite danger language.

At present, several MT approaches such as rule-based machine translation (RBMT), statistical machine translation (SMT), NMT, and transformer model are available. Many researchers consider transformer model as a state-of-the-art model in the automatic translation world as it provides a high accuracy rate compared to all other approaches. In this paper, by the advantages of less training time and better translation accuracy in complex data, we attempt to use SMT using MOSES. MOSES is a machine translation package (toolkit) that is user-friendly and easily understandable by any user and finally evaluates the translation accuracy and adequacy of the corpus. We have used the BLEU score to evaluate the quality of translation [1, 2].

2 Related Works

The Nyishi language lacks research work in MT since this language doesn't have record of any previous documentation. We choose to use phrase-based MT method for the low-resourced Nyishi Language as this approach helps in meaning conservation rather than the structural conservation. The resource we used is manually translated as it is low resource, and we tend to analyze the data with the model whose priority is to recommend the meaning of the language. Several research works have been done on different low-resource languages like English–Hindi, Hindi–Nepali, English–Tamil, English–Mizo (Amarnat et al. 2018), English–Punjabi, Hindi–Nepali, and Punjabi–Hindi by using SMT; related works provide the better translation accuracy. SMT overcomes the problem of the rule-based system by supporting the n-gram model, where basic steps are finding the maximum probabilities of phrase pairs in parallel source–target sentences. Finally, BLEU is used to calculate the automatic metrics of the predicted translation [3–6].

2.1 SMT for Low-Resource Language

Regardless of the word length, phrase-based MT reduces the restriction of word-based translation by translating the whole sequence of words. The sequence of the words is called blocks or phrases. Among several machine translation approaches, we chose to use SMT because it doesn't depend on language attributes. This model supports training data on the small and large dataset as well as helps in preserving meaning of the language better; increasing of dataset will eventually provide better translation results. The SMT uses the Bayes theorem for probability distribution [7].

Bayes rule:

$$\begin{aligned} \tilde{e} &= \arg\max_{e} \frac{p_{TM}(n|e)\, p_{LM}(e)}{p(n)} \\ &= \arg\max_{e} p_{TM}(n|e)\, p_{LM}(e) \end{aligned}$$

where

- Translation model $p_{TM}(n|e)$ is the probability where the source string is the translation of the target string.
- Language model $p_{LM(e)}$ is the probability used for distinguishing that target language string.
- Where $\tilde{e}$ is the best translation concluded by choosing the one that provide highest probability?

2.2 Moses in PBSMT

Phrase based and tree based are the two translation models that come with Moses toolkit**;** in this paper, we use phrase-based approaches to obtain the translation accuracy on the parallel corpus. Moses supports several languages modeling toolkits, such as SRILM, KenLM, IRSTLM, and RandLM, and also allowed to introduce a new toolkit. GIZA supports word aligning for given parallel corpus and language model estimation. The decoder of Moses has several procedures such as chart parsing and cube pruning. In our paper, we use MGIZA for alignment, cube pruning for identifying the number of hypotheses to be covered; IRSTLM and RandLM are used for language modeling.

3 System Descriptions

This paper analyzes the translation accuracy by using the SMT with phrase-based approaches, and other parts of the paper are as follows: corpus preparations, language

model, translation model, decoding, experiment part, and results, which are evaluated by BLEU, NIST; and finally, human evaluation is done to calculate the overall translation accuracy and adequacy.

3.1 Corpus Preparations

The parallel aligned sentence pair of 30,000 with both backward and forward direction has been developed, where English is the source and Nyishi is the target language and vice versa. Due to low resources and lack of script in this language, the preparation of corpus is complex and uncertain. The collected parallel sentences of Nyishi have been verified thoroughly by a native of Nyishi language experts. In PBSMT, corpus preparation includes tokenization, true casing, and cleaning. True casing helps to handle the difference between uppercase and lowercase words and finally convert all to lowercase forms. The cleaning step drops down the long and unwanted sentence along with the threshold exceeding chunk.

3.2 Language Model

Language models in SMT reorder the words/phrases that are suggested by the translation model to generate the target language. Sentence probability is calculated by the probability method in the SMT language model, using the n-gram model. Computation of the probability in the language model is based on single word that provides all the words that lead to make a whole sentence [8]. IRSTLM tool is used to develop the language model. Language modeling in PBSMT helps to break the probability of a sentence $P(S)$ with the probability of individual words $P(W)$ shown in the below equation:

$$\begin{aligned} P(S) &= P(W1, W2, W3, \ldots, Wn \\ &= P(W1)P(W2|W1)P(W3, |W1W2) \\ &\quad P(W4|W1W2W3) \ldots P(Wn|W1W2 \ldots Wn-1) \end{aligned}$$

Initially, the probability of a word in a sentence is calculated following which the probability of sentences is calculated, which gives the orders of word proceeding to it and n-gram model is used to generate the probability approximation of all the previous words in the sentences [5].

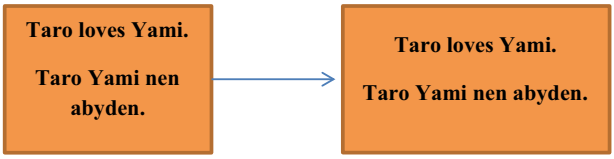

Fig. 1 Word alignment with MGIZA

3.3 Translation Model

The main objective of the translation model is to advise a set of possible words/phrases for a given source sentence and compare the expression of meaning between languages. This model is known as the translation model because of its expression comparing properties between different languages. It shows that a large quantity of the corpus will produce a high translation score than the low resources dataset as MT supports the better result of translation accuracy with a large dataset. Alignment model building between the input and the output language is the first step of translation models [9]. In every sentence of the training parallel corpus, the job of providing the best set of alignment links is handled by the alignment models. The MGIZA in Moses toolkit has been used to align our Nyishi–English and vice versa training corpus. The word-align text in alignment model in SMT is necessary, as it helps to initialize the translation model in machine translation. Such as, finding words that correspond to each other performed automatically with the probabilistic method as elaborated below [10] (Fig. 1):

$$P(\text{Yami}|\text{Yami}) = 0.99, \quad P(\text{Taro}|\text{Taro}) = 0.97.$$
$$P(\text{loves}|\text{abyden}) = 0.46, \quad P(\text{loves}|\text{nen}) = 0.04.$$

3.4 Decoding

The main function of the decoder is to identify the better candidate translation and to search for the hypothesis which contains the best model score. The output of the LM, TM, and the input for the decoder are the source sentence. In the decoding process, computational complexity is high. We used the beam search approach for searching strategy in the decoder which supports heuristic-based algorithm strategies. Decoder generally uses mathematical approach of most probable translation, for which probability of an individual-targeted word/phrase is maximum. Figure 2 shows the PBSMT system conceptual architecture [5].

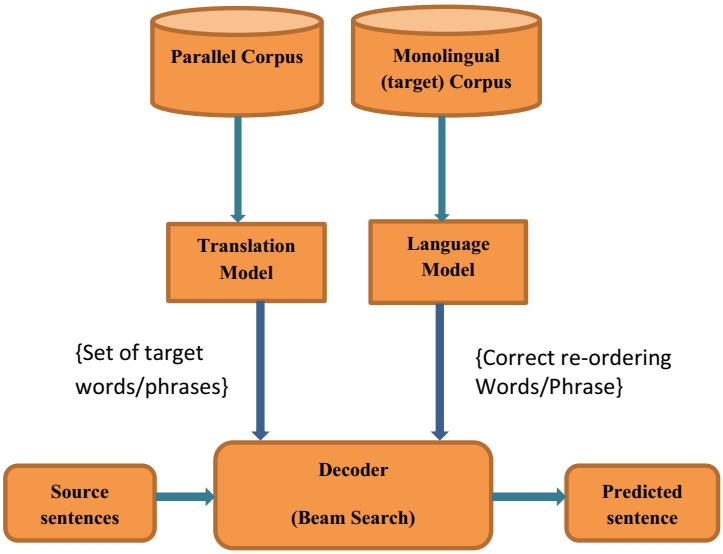

Fig. 2 Abstract architecture of PBSMT system [5]

3.5 PBSMT Experimental Design for English–Nyishi

The newly collected parallel aligned corpus of 30,000 of the English–Nyishi is divided into three parts as training, tuning, and testing data, and English sentence of the monolingual file has been used for building the language model. The language model is built with the IRSTLM toolkit in the Moses that is used to assist data structures and algorithm, where language models are suited for accessing and storing large n-gram data. Data used are shown as in Table 1.

4 Results and Analysis

Results that are generated from the SMT systems have been calculated by the BLEU metric, and human evaluation has also been performed for the same.

Table 1 Corpus statistics

Corpus type	English–Nyishi	Nyishi–English
Training data	28,033	28,033
Tuning data	967	967
Testing data	1001	1001

4.1 About BLEU and NIST Score

BLEU used in MT is an algorithm that attempts to provide the accuracy of the text translation by a machine-translated output. The output of the BLEU is always between 0 and 1, whereas NIST ranges from 1 to 10. BLEU/ NIST metrics only depend on precision. The score of BLEU/NIST is calculated from n-grams precision [25]. Test set is required for BLEU/NIST metrics machine translation system. The equation of the BLEU score for n-gram is calculated by using the following formula:

$$\min\left(1, \frac{\text{candidate} - \text{length}}{\text{reference} - \text{length}}\right)\left(\prod_{i=1}^{n} \text{precision}_i\right)^{\frac{1}{n}}$$

Here, candidate-length stands for candidate translation length, and the reference-length stands for reference translation length, and also precision I denotes precision score [25] for *i*th gram match. NIST and BLEU identify the translation accuracy with tuning and without tuning which clearly shows that with tuning in both the translation accuracy is better than without tuning that is because tuning maximizes the performance of translation on a small set of parallel sentences by finding the optimal weights for this linear model (Tables 2 and 3).

The individual n-gram score generated from the model shows that results are comparatively better in unigram which means translation accuracy is high in short sentences compare to medium and long sentences (Table 4).

The score generated by English–Nyishi and vice versa is low compared to other high-resources language due to the high ambiguity of Nyishi words.

Table 2 Individual n-gram by BLEU score

No.	Corpus type	BLEU score without tuning			BLEU score with tuning		
	Individual n-gram	1-g	2-g	3-g	1-g	2-g	3-g
1	English–Nyishi	0.3624	0.1357	0.0744	0.3798	0.1792	0.1184
2	Nyishi–English	0.3497	0.1479	0.0988	0.3699	0.1937	0.1402

Table 3 Individual n-gram by NIST score

No.	Corpus type	NIST score without tuning			NIST score with tuning		
	Individual n-gram	1-g	2-g	3-g	1-g	2-g	3-g
1	English–Nyishi	1.6744	0.2717	0.0213	2.3398	0.4510	0.0534
2	Nyishi–English	2.2148	0.3617	0.0571	2.3660	0.4861	0.0820

Table 4 N-gram scoring English–Nyishi

No.	Corpus type	NIST score without tuning	BLEU score without tuning	NIST score with tuning	BLEU score with tuning
1	English–Nyishi	1.9700	0.0802	2.8512	0.1411
2	Nyishi–English	2.6408	0.1419	2.9444	0.1849

4.2 Human Evaluation

Human evaluation evaluates the adequacy and fluency as well as the overall rating of translated results. Human evaluation is necessary for MT as BLEU scores fail to meet the quality of translated score in terms of adequacy and fluency; these two factors help to identify the translation quality of MT. Adequacy helps to identify the total number of similar meanings between reference translation and candidate translation. Fluency is measured by the translated sentence of reference sentence irrespective of candidate sentence, where fluency is considered when reference sentence is translated fluently. As the BLEU fail to enclose all the features of evaluating the candidate translation. The output of quality translation is measured with the reference translation; the basic criteria of measurements are totally on fluency and adequacy. In this experiment, Nyishi language expert is from (NBCC) Nyishi Baptist church council and a native of Nyishi, who have a great knowledge of the Nyishi language and also a core in charge of Nyishi language development in all Nyishi elite society and other who contributed in translation. The core of the Nyishi language and where the translation is to be used is well conscious of the experts; however, the human evaluation process is very expensive, time-consuming.

5 Analysis of Translation

To analyze the translation quality from the output of the SMT system in both directions, we have used few sample sentences from translated output where the quality is judged from various issues like under-translation, over-translation, and wrong translation of name entities. Therefore, the selected translated sentence has been judged on two factors, i.e., adequacy and fluency where system translations are examined in case of best or worst. Although corpus is analyzed in both the direction but most common error that we encounter are mixed translation of source word and target word, **English Gold (EG)** in English is the reference sentence, **Nyishi Test (NT)** as a test sentence, **and English Predicted (EP)** in English the predicted sentence.

1. Best performance of translation which is adequate and fluent.
   NT: “Svka”
   EG: “Assist”
   EP: “Help”

2. The system has predicted the unigram and tri-gram correctly "I, him," and therefore, it is not *adequate* but is *fluent.*
   NT: "Ngo mwvn kapapa"
   EG: "I saw him"
   EP: "I met him"
3. Prediction result with *partially adequate* and *perfectly fluent* translation.
   NT: "Nyem ko"
   EG: "daughter"
   EP: "Girl"
4. Prediction result predicted as neither *fluent* nor *adequate.*
   NT: "Tom danypa"
   EG: "Tom is simple"
   EP: "Tom is ordinary"
5. Prediction result that is *adequate* as well as *fluent* and automatically adds "7" which is numerically predicted from sentences.
   NT: "Hv sija kvn paku"
   EG: "It is seven now"
   EP: "Now it is 7"
6. The SMT system has made a prediction that mixed word of target and source to the predicted sentence.
   NT: "No Tom nen svka numyv?"
   EG: "Did you help Tom?"
   EP: "You numyv let tom help?"

From the above analysis of the translations predicted by the PBSMT system, human evaluation is done to evaluate the translated results in factors of adequacy and fluency. Further, it shows that score degrades as the sentence length increases, and scores are comparatively better in short and medium sentences than of long sentences.

6 Conclusion and Future Work

MT is a communication tool to bridge among the people belonging to the different linguistic backgrounds around the world. This work attempts to introduce the low-resource language Nyishi into a machine translation environment with few corpora, and we successfully analyzed by using SMT; furthermore, we intend to increase our corpus to get better prediction result as it requires large corpus to provide better translation using MT methods. To minimize the error, we aim to identify more statistical approaches as well as the neural method to improve adequacy and fluency.

References

1. Dey, M.: Negation in Nyishi. NEHU J. **15**(2), 79–100 (2017)
2. John, S.S., Lomdak, L.: Language Endangerment in Arunachal Pradesh: Current Issues and Future Prospects. Centre for Endangered Languages (CFEL) Rajiv Gandhi University (2017)
3. Laskar, S.R., Dutta, A., Pakray, P., Bandyopadhyay, S.: Neural machine translation: English to Hindi. In: IEEE Conference on Information and Communication Technology, pp. 1–6 (2019)
4. Laskar, S.R., Dutta, A., Pakray, P., Bandyopadhyay, S.: Neural Machine Translation: Hindi⇔Nepali. In: Proceedings of the Fourth Conference on Machine Translation (WMT), Vol. 3 Shared Task Papers, pp. 202–207. Association for Computational Linguistics, Italy (2019)
5. Pathak, A., Pakray, P., Bentham, J.: English-Mizo Machine Translation using neural and statistical approaches. Neural Comput. Appl. **31**(11), 7615–7631 (2019)
6. Rushanti, K., Sambyo, K.: Phrase-based machine translation of Digaru-English. In: Electronic Systems and Intelligent Computing, pp. 983–992. Springer (2020)
7. Graham Neubig Homepage, http://www.phontron.com/slides/building-smt-en-20120510.pdf. Last accessed 2021/03/11
8. Baruah, K.K., Das, P., Hannan, A., Sarma, S.K.: Assamese-English Bilingual Machine Translation. Int. J. Nat. Lang. Comput. (IJNLC) **3**(30), 73–82 (2014)
9. Zin, T.T., Soe, K.M., Thein, N.L.: Translation model of Myanmar phrases for statistical machine translation. In: International Conference on Intelligent Computing, pp. 235–242. Springer, Heidelberg (2011)
10. Hu, Y.: Statistical machine translation based on translation rules. J. Chem. Pharm. Res. **6**(7), 1628–1635 (2014)
11. Romdhane, A.B., Jamoussi, S., Hamadou, A.B., Smaïli, K.: Phrase-based language model in statistical machine translation. Int. J. Comput. Linguist. Appl. (2016)
12. Raghavendra, U.U., Tanveer, A.F.: An English-Hindi statistical machine translation system. In: International Conference on Natural Language Processing 2004, IJCNLP, pp. 254–262. Natural Language Processing (2004)
13. Cho, K., Merriënboer, B.V., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: encoder-decoder approaches. In: Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, pp. 103–111. Association for Computational Linguistics, Doha (2014)
14. Kishore, P., 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, pp. 311–318. Association for Computational Linguistics (2002)
15. Denkowski, M., Lavie, A.: Meteor universal: language specific translation evaluation for any target language. In: Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 376–380. Association for Computational Linguistics, Baltimore (2014)
16. Koehn, P., Hoang, H., Birch, A., Burch, C.C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C.: Moses: open source toolkit for statistical machine translation. In: Proceedings of the ACL 2007 Demo and Poster Sessions, pp. 177–180. Association for Computational Linguistics, Prague (2007)
17. Dugonik, J., Boskovic, B., Maucec, M.S., Brest, J.: The usage of differential evolution in a statistical machine translation. In: IEEE Symposium on Differential Evolution (SDE), pp. 1–8. IEEE, Orlando (2014)
18. Nabhan, A.R., Rafea, A.: Tuning statistical machine translation parameters using Perplexity. In: IRI—IEEE International Conference on Information Reuse and Integration, IEEE, Las Vegas (2005)
19. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566–4575. IEEE, Boston (2015)
20. Ganguly, D., Leveling, J., Jones, G.J.F.: Bengali (Bangla) information retrieval. Technical Challenges and Design Issues in Bangla Language Processing, IGI Global (2013)

21. Faili, H.: An experiment of word sense disambiguation in a machine translation system. In: International Conference on Natural Language Processing and Knowledge Engineering, pp. 1–7. IEEE, Beijing (2008)
22. Sudhahar, S., Cristianini, N.: Detecting shifts in public opinion: a big data study of global news content. In: International Symposium on Intelligent Data Analysis, pp. 316–327. Springer Cham (2018)