NUIG-Panlingua-KMI Hindi ↔Marathi MT Systems for Similar Language Translation Task @ WMT 2020

Atul Kr. Ojha[!+], Priya Rani[!], Akanksha Bansal[+], Bharathi Raja Chakravarthi[!], Ritesh Kumar[*], John P. McCrae[!]

[!]Data Science Institute, NUIG, Galway, [+]Panlingua Language Processing LLP, New Delhi, [*]Dr. Bhimrao Ambedkar University, Agra

(atulkumar.ojha,priya.rani,bharathi.raja)@insight-centre.org, panlingua@outlook.com, john.mccrae@nuigalway.ie, ritesh78_llh@jnu.ac.in

Abstract

NUIG-Panlingua-KMI submission to WMT 2020 seeks to push the state-of-the-art in the Similar language translation task for the Hindi ↔ Marathi language pair. As part of these efforts, we conducted a series of experiments to address the challenges for translation between similar languages. Among the 4 MT systems prepared for this task, 1 PBSMT systems were prepared for Hindi ↔ Marathi each and 1 NMT systems were developed for Hindi ↔ Marathi using Byte Pair Encoding (BPE) of subwords. The results show that different architectures in NMT could be an effective method for developing MT systems for closely related languages. Our Hindi-Marathi NMT system was ranked 8th among the 14 teams that participated and our Marathi-Hindi NMT system was ranked 8th among the 11 teams participated for the task.

1 Introduction

Developing automated relations between closely related languages is a contemporary concern especially in the domain of Machine Translation(MT). Hindi and Marathi exhibit a significant overlap in their vocabularies and strong syntactic plus lexical similarities. These striking similarities seem promising in enhancing the possibility of mutual inter-comprehension within closely related languages. However, automated translation between such closely related languages is a rather challenging task.

The linguistic similarities and regularities in morphological variations and orthography motivate the use of character-level translation models, which have been applied to translation (Vilar et al., 2007; Chakravarthi et al., 2020) and transliteration (Matthews, 2007; Chakravarthi et al., 2019a; Chakravarthi, 2020). In the past few years, neural machine translation systems have achieved outstanding performance with high resource languages, with the help of open source toolkit such as OpenNMT (Klein et al., 2017), Marian (Junczys-Dowmunt et al., 2018) and Neamtus (Sennrich et al., 2017), which provide various ways of experimenting with the use of different features and architectures, yet it fails to achieve the same results with low resource languages (Chakravarthi et al., 2018, 2019b). However, Sennrich and Zhang (2019) revisited the NMT models and tuned hyper-parameters, changed network architectures to optimize NMT for low-resource conditions and concluded that low-resource NMT is very sensitive to hyper-parameters such as Byte Pair Encoding (BPE) vocabulary size, word dropout, and others. This paper is an extension of our work Ojha et al. (2019) submitted to WMT 2019 similar language translation task. Therefore our team adapted methods of the low resource setting for NMT proposed by Sennrich and Zhang (2019) to explore the following broad objectives:

- to compare the performance of SMT and NMT in case of closely related, relatively low-resourced language pairs, and
- to findout how to leverage the accuracy of NMT in closely related languages using BPE into subwords.
- to analyze the effects of data quality in performance of the systems.

2 System Description

This section provides an overview of the systems developed for the WMT 2020 Shared Task. In these experiments, the NUIG-Panlingua-KMI team explored two different approaches: phrase-based statistical (Koehn et al., 2003), and neural method for Hindi-Marathi and Marathi-Hindi language pairs. In all the submitted systems, we use the Moses (Koehn et al., 2007) and Nematus (Sennrich et al., 2017) toolkit for developing statistical and neural

Proceedings of the 5th Conference on Machine Translation (WMT), pages 418–423
Online, November 19–20, 2020.

machine translation systems respectively. The pre-processing was done to handle noise in data (for example, different language sentences, non-UTF characters etc), the details of which are provided in section 3.1

2.1 Phrase-based SMT Systems

These systems were built on the Moses open source toolkit using the KenLM (Heafield, 2011) language model and GIZA++ (Och and Ney, 2003) aligner. 'Grow-diag-final-and heuristic' parameters were used to extract phrases from the corresponding parallel corpora. In addition to this, KenLM was used to build 5-gram language models.

2.2 Neural Machine Translation System

Nematus was used to build 2 NMT systems. As we mentioned in an earlier section, at first data was pre-processed at subwords level with BPE for neural translation, and then the system was trained using Nematus toolkit. Most of the system features were adopted from (Sennrich et al., 2017; Koehn and Knowles, 2017) (see section 3.3.2).

2.3 Assessment

Assessment of these systems was done on the standard automatic evaluation metrics: BLEU (Papineni et al., 2002), Rank-based Intuitive Bilingual Evaluation Score (RIBES) (Isozaki et al., 2010) and Translation Error Rate (TER) (Snover et al., 2006).

3 Experiments

This section briefly describes the experiment settings for developing the systems.

3.1 Data Preparations

The parallel data-set for these experiments was provided by the *WMT Similar Translation Shared Task* [1] organisers and the Marathi monolingual data-set was taken from *WMT 2020 Shared Task: Parallel Corpus Filtering for Low-Resource Conditions*.[2] The parallel data was sub-divided into training, tuning, and monolingual sets, as detailed in Table 1. However, the shared data was very noisy.

To enhance the data quality, the team had to undertake an extensive pre-processing session focused on identifying and cleaning the data-sets. Out of 43274 training sentences, the Hindi corpus had Telugu sentences while the Marathi corpus had Meitei sentences intermingled as shown in first row (Figure 1). The parallel data had more than 1192 lines that were not comparable with each other as shown in second and third row (Figure 1), where some Hindi sentences had only half the sentences translated in Marathi (second row) and some had blank spaces against their Marathi counter parts (third row). The translation quality of the parallel data was also not up to mark. In fact, the team could locate a few instances of synthetic data. There were a few sentences where character encoding was an issue, hence were completely unintelligible.

Language Pair	Training	Tuning	Monolingual
Hindi ↔ Marathi	43274	1411	-
Marathi	-	-	326748
Hindi	-	-	75348193

Table 1: Statistics of Parallel and Monolingual Sentences of the Hindi and Marathi Languages

3.2 Pre-processing

The following pre-processing steps were performed as part of the experiments:

a) Both corpora were tokenized and cleaned (sentences of length over 80 words were removed).

b) For neural translation, training, validation and test data was prepossessed into subwords BPE format. This format was utilised to prepare BPE and vocabulary further used.

All these processes were performed using Moses scripts. However, the tokenization was done by the RGNLP team tokenizer (Ojha et al., 2018) and Indic_nlp_library.[3] These tokenizers were used since Moses does not provide a tokenizer for Indic languages. Also the RGNLP tokenizer ensured that the canonical Unicode representation of the characters are retained.

3.3 Development of the NUIG-Panlingua-KMI MT Systems

After removing noisy and pre-processing data, the following steps were followed to build the NUIG-Panlingua-KMI MT systems:

[1] http://www.statmt.org/wmt20/similar.html

[2] https://wmt20similar.cs.upc.edu/

[3] https://github.com/anoopkunchukuttan/indic_nlp_library

Sr.No	Hindi Sentences	Marathi Sentences
1	ಬಾಳೆಯಣ್ಣು , ಪೊಪ್ಪಾಯ , ಸೀತಾಫಲ ಮುಂತಾದವುಗಳು ಶೀಣ ಗುಣ ಹೊಂದಿದ್ದರೆ ಶುಂಟಿ , ಬೆಳ್ಳುಳ್ಳಿ , ಈರುಳ್ಳಿ ಮುಂತಾದವು ಉಷ್ಣ ಗುಣವನ್ನು ಹೊಂದಿರ	[illegible]
2	सचिन तेंदुलकर ने एकदिवसी क्रिकेट में सर्वाधिक शतक बनाकर , क्रिकेट जगत में अपनी प्रधानता सिद्ध कर दी । / आधुनिक युग में वैज्ञानिकों की प्रधानता को झुठलाया	आजच्या शिक्षणपद्धतीत विज्ञान विषयाला प्राधान्य आहे ।
3	अप्रियभाषी व्यक्ति निंदा का पात्र होता है। / अप्रियभाषी व्यक्ति अपनी बोली से अपनों में भी पराया बना रहता है ।	-
4	हर देशवासी को हमारी इस मातृ-शक्ति पर, हमारी इन वीरांगनाओं के प्रति आदर होना बहुत स्वाभाविक है। मैं इन दोनों बहनों को हृदय से बहुत-बहुत बधाई देता हू।	प्रत्येक देशवासियाला आपल्या या मातृ-शक्ती विषयी, आपल्या या वीरांगनांविषयी आदर वाटणे स्वाभाविक आहे.

Figure 1: Examples of discrepancies in Hindi-Marathi parallel data

Example-1	
Source	अपने संबोधन के दौरान, प्रधानमंत्री ने समारोह में उपस्थित व्यक्तियों को, विशेषकर के जो पहले भारत नहीं आ चुके हैं, उन्हें भारत की सांस्कृतिक एवं आध्यात्मिक परंपरा का अनुभव करने के लिये जनवरी 2019 में आयोजित होने वाले कुंभ मेले में आने के लिये आमंत्रित भी किया।
Reference	आपल्या भाषणादरम्यान, पंतप्रधानांनी भारतातील सांस्कृतिक आणि आध्यात्मिक परंपरेचा अनुभव घेण्यासाठी जानेवारी 201 9 मध्ये आयोजित कुंभमेळ्याच्या निमित्ताने भारताला अजून पर्यंत भेट न दिलेल्या लोकांना उपस्थित राहण्याचे आवाहन केले.
NUIG-Panlingua-KMI_P	यावेळी दौरान, पंतप्रधानांनी समारंभात उपस्थित व्यक्ती को, विशेषत : जे आधी भारत नाही आले हैं, त्यांना भारतातील सांस्कृतिक आणि आध्यात्मिक परंपरेचा अनुभव करण्यासाठी जानेवारी 2019 मधील होणाऱ्या कुंभमेळ्यात येण्यासाठी निमंत्रण दिले .
NUIG-Panlingua-KMI_C	बोलतांना दौरान, पंतप्रधानांनी दिलेल्या को, विशेषत : जे आधी भारत नाही आलो हैं, भारतातील सांस्कृतिक आणि आध्यात्मिक परंपरेचा अनुभव घेण्यासाठी जानेवारी 201 9 मध्ये आयोजित कुंभमेळ्याच्या निमित्ताने भारताला अजून पर्यंत भेट न .
Example-2	
Source	आमचा राष्ट्राला अर्पण केलेला पच्छिम मालवाहतूक कॉरिडॉर आता पूर्ण होणार आहे.
Reference	हमारा वेस्टर्न डेडिकेटेड फ्रेट कॉरिडोर अब पूरा होने वाला है।
NUIG-Panlingua-KMI_P	हमारा राष्ट्र को समर्पित किया पच्छिम कार्गो कॉरिडोर अब पूरी हो गई है।
NUIG-Panlingua-KMI_C	हमारा राष्ट्र को समर्पित करने के पच्छिम माल कॉरिडोर अब लगभग पूरा होने वाला है ।
Example-3	
Source	उन्होंने शौचालय की प्रतिष्ठा बढ़ा दी है।
Reference	त्यांनी शौचालयाची प्रतिष्ठा वाढवली आहे.
NUIG-Panlingua-KMI_P	त्यांनी शौचालय यांच्या प्रतिष्ठा वाढवली आहे .
NUIG-Panlingua-KMI_C	त्यांनी शौचालयांना प्रतिष्ठा वाढवली आहे .

Figure 2: Analysis of the PBSMT and NMT's Systems

3.3.1 Building Primary MT Systems:

As previously mentioned, the Hindi-Marathi and Marathi-Hindi PBSMT systems were built as the primary submission using Moses. The language model was built first, using KenLM. For Marathi-Hindi and Hindi-Marathi language pairs, the language models were trained on 5-gram. After that, the systems were built independently and combined in a loglinear scheme in which each model was assigned a different weight using the Minimum Error Rate Training (Och, 2003) tuning algorithm. To train and tune the systems, we used 40454 and 1411

parallel sentences, respectively, for all language pairs.

3.3.2 Building Contrastive MT Systems:

As mentioned in the previous section, Nematus toolkit was used to develop the NMT systems. The training was done on subword and character-level. All the NMT experiments were carried out only with a data-set that contained sentences with length of up to 80 words. The neural model is trained on 5000 epochs, using Adam with a default learning rate of 0.002, dropout at 0.01 and mini-batches of 80 and the batch size for the validation was 40. Vocabulary size of 30000 for both Marathi-Hindi and Hindi-Marathi language pairs was extracted. Remaining parameters were limited with the use of default hyper-parameters configuration.

4 Evaluation

All the systems were evaluated using the reference set provided by the shared task organizers. The standard MT evaluation metrics, BLEU (Papineni et al., 2002) score, RIBES (Isozaki et al., 2010) and TER (Snover et al., 2006), were used for automatic evaluation. These results were prepared on the Primary and Contrastive system submission which are mentioned in the Table 2 as _P and _C, where *_P* stands for Primary and *_C* stands for Contrastive, respectively. It gives a quantitative picture of particular differences across different systems, especially with reference to evaluation scores (Table 2)

System	BLEU	RIBES	TER
Hindi-Marathi_P	9.38	51.88	91.24
Hindi-Marathi_C	9.76	52.18	91.49
Marathi-Hindi_P	17.38	59.31	81.47
Marathi-Hindi_C	17.39	58.84	81.15

Table 2: Accuracy of Hindi↔Marathi MT Systems at BLEU, RIBES and TER Metrics

4.1 Results

Overall we see varying performance among the system submitted to the task, with some performing much better out-of-sample than others. The NUIG-Panlingua-KMI subword NMT system took 8th position for both Hindi-Marathi and Marathi-Hindi language pair, across 14 teams. Our subword NMT systems for Marathi-Hindi language pair showed better results in terms of all the three metrics (17.39 in BLEU, 58.84 in RIBES and 81.15 in TER) while the Hindi-Marathi language pair scored 9.76 in BLEU, 52.18 in RIBES and 91.24 in TER. Across both the language pairs, subword based NMT performed better than PBSMT as its accuracy rate was higher in BLEU and lower in TER metrics, shown in Table 2.

4.2 Analysis

We used the reference set provided by the shared task organizers to evaluate both PBSMT and NMT systems. Even though subword based NMT system could take advantage of the shared features among similar languages, challenges in translating a few linguistics structures acted as a constraint. Example 1 shown in Figure 2 is one of the challenging structures that the system was unable to translate. In these sentences the systems could not capture the correct tense and aspect which is past perfect in source sentence whereas the NMT system translated it as simple past. The second most common challenging structures that needed special attention were the postpositions as shown in Example 2 and 3 in the figure. In most cases, the system over-generalised the sentences in Marathi and generated unnecessary postposition phrases in Hindi as in Example 2. Similarly, we can see in Example 3 while translating from Hindi to Marathi both PBSMT and NMT systems used wrong post-positions.

5 Conclusion

Our experiment results reveal that subword based NMT could take advantage of the relation between the similar language to boost the accuracy of neural machine translations system in low resource data settings. As BPE units are variable-length units and the vocabularies used are much smaller than morpheme and word-level model, the problem of data sparsity does not occur. On the contrary, it provides an appropriate context for translation between similar languages. However, the quality of data used to train the systems does affect the quality of translation. Thus, we could conclude that shared features between two languages could be an advantage to leverage the accuracy of NMT systems for closely related languages.

Acknowledgments

This publication has emanated from research in part supported by the Irish Research Council under grant number SFI/18/CRT/6223 (CRT-Centre for Research Training in Artificial Intelligence) co-funded by the European Regional Development Fund as well as by the EU H2020 programme un-

der grant agreements 731015 (ELEXIS-European Lexical Infrastructure).
We are also grateful to the organizers of WMT Similar Translation Shared Task 2020 for providing us the Hindi↔Marathi Parallel Corpus, monolingual and evaluation scores.

References

Bharathi Raja Chakravarthi. 2020. *Leveraging orthographic information to improve machine translation of under-resourced languages*. Ph.D. thesis, NUI Galway.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P McCrae. 2018. Improving wordnets for under-resourced languages using machine translation. In *Proceedings of the 9th Global WordNet Conference (GWC 2018)*, page 78.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P McCrae. 2019a. Comparison of different orthographies for machine translation of under-resourced Dravidian languages. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Bharathi Raja Chakravarthi, Mihael Arcan, and John Philip McCrae. 2019b. Wordnet gloss translation for under-resourced languages using multilingual neural machine translation. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*, pages 1–7.

Bharathi Raja Chakravarthi, Priya Rani, Mihael Arcan, and John P McCrae. 2020. A survey of orthographic information in machine translation. *arXiv e-prints*, pages arXiv–2008.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197. Association for Computational Linguistics.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. 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*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

David Matthews. 2007. Machine transliteration of proper names. *Master's Thesis, University of Edinburgh, Edinburgh, United Kingdom*.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Atul Kr Ojha, Koel Dutta Chowdhury, Chao-Hong Liu, and Karan Saxena. 2018. The RGNLP machine translation systems for WAT 2018. In *Proceedings of the 5th Workshop on Asian Translation (WAT2018)*.

Atul Kr Ojha, Ritesh Kumar, Akanksha Bansal, and Priya Rani. 2019. Panlingua-KMI MT system for similar language translation task at WMT 2019. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 213–218.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, et al. 2017. Nematus: a toolkit for neural machine translation. *arXiv preprint arXiv:1703.04357*.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200.

David Vilar, Jan-Thorsten Peter, and Hermann Ney. 2007. Can we translate letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 33–39, Prague, Czech Republic. Association for Computational Linguistics.