Leveraging related languages for the translation of a low-resourced language pair: Manipuri to English

Rudali HUIDROM and Yves LEPAGE

Graduate School of Information, Production and Systems
Waseda University

{rudali.huidrom@ruri.,yves.lepage@}waseda.jp

13th November 2020

Outline

Language Description

Manipuri (locally known as, Meiteilon), is a low resourced Indian language:

- **ISO 639-2**: *mni.*
- **Language Family**: *Sino-Tibetan.*
- **Writing system**: *Eastern Nagari script* *(also known as, Bengali Script) and Meitei Mayek.*
- **Same writing system**: *Manipuri, Bengali*[1] *and Assamese*[2].
- **Lexical influence**: *Bengali influences Assamese and Manipuri. Hindi*[3] *influences all the other three languages.*

ISO 639-2 code: [1] *ben*, [2] *asm*, [3] *hin*.

Dataset Description

Source: *pmindia dataset* (Haddow et al., 2020)

Language pair	sentence pairs	words / sent.	word types
Assamese	9,732	17	26,649
English		20	22,900
Bengali	29,584	15	55,150
English		17	38,781
Hindi	56,831	20	52,441
English		19	59,061
Manipuri	7,419	15	22,289
English		19	18,502

Table1: Statistics on the dataset used.

Motivation

Our main goal: *To increase the translation accuracy of Manipuri to English.*

Proposal: *To build different machine translation systems by adding knowledge from the three languages in combination with Manipuri and learn translation models.*

Our method

1. **Baseline:**

   Translation of Manipuri to English.

Our method

1. **Baseline:**

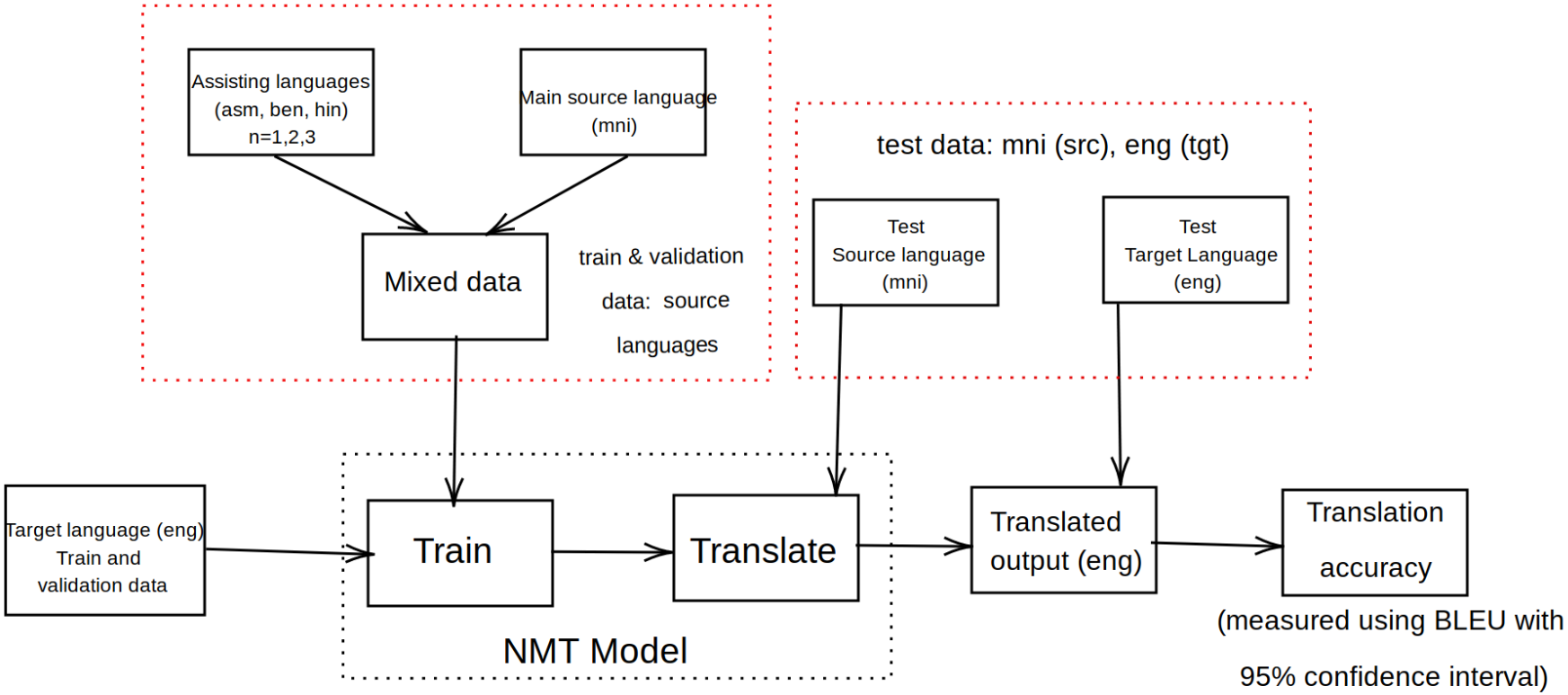
   Translation of Manipuri to English.

2. **Combination with other language(s):**

   *Translation by combining Manipuri with **'n'** number of languages to English; n = 1,2,3.*

Explanation

Figure 1. Explanation of the working of our method. Translation by combining Manipuri with **'n'** number of languages to English; n = 1,2,3.



Experiments

Languages used:

Source languages: *Manipuri (mni)*, *Bengali (ben)*, *Assamese (asm)*, *Hindi (hin)*.

Target language: *English (eng)*.

Experiment setup:

Tool settings: *OpenNMT (Klein et al., 2017) RNN model (encoder type: brnn; decoder type: rnn); dropout rate: 0.3.*

Data settings: *5000 training data for every language pair and 1000 data each for validation and test for Manipuri-English language pairs.*

Result

Manipuri	Assamese	Bengali	Hindi	BLEU	Confidence Interval
■				4.3	±0.6
■	■			6.7	±0.8
■		■		5.9	±0.7
■			■	6.6	±0.7
■	■	■		10.5	±1.1
■	■		■	10.1	±1.1
■		■	■	9.4	±0.9
■	■	■	■	12.7	±1.3

Table2: Translation accuracies as measured with BLEU with 95% confidence.

Conclusion

- **Conclusion**: *Lexical influence exhibited by related languages helped in increasing the translation accuracy of Manipuri to English.*

  A linear increase of translation accuracy is observed, independently of languages.

- **Future work**: *To conduct the experiment with more amount of data such as, data from our corpora.*

Thank you for listening!

Experiment settings

RNN model	
Embed Dim	500
RNN Type	LSTM
Num Layers	2
Hidden Dim	500
Input Feeding	True
Attention	Global
Attention type	General
Dropout	0.3
Encoder Type	brnn
Decoder Type	rnn
Optimization	
Batch size	64
Batch type	Sentences
Optimizer	adam
Init learning rate	0.001
Learning rate schedule	
# steps before decay	50,000
Decay frequency	10,000
Learning rate decay	lcurr * 0.5

Table3: Parameters used for RNN model. They are mostly from openNMT-py toolkit suggestions.

Background

- Languages are often not isolated (Song et al., 2020), it has lexical influences on one other which results from either belonging to the same language family or the regional influences on the languages.
- We are ***observing the phenomenon*** of leveraging translation knowledge from other languages.
- The notion of leveraging related languages for translation is widely applied in research areas of MT as well as in commercial fields[4].

[4] *https://www.sdl.com/*

Background

- Statistical MT is based on ***purely bilingual systems***; in particular, it is a bilingual model.
- On the other hand, Neural MT's ***shared representation space*** is forced across languages and it induces a sort of transfer-learning (Lakew et al., 2018) which helps in increasing the translation accuracy[5].

[5] *https://research.google/pubs/pub49064/*

Bibliography

1. Barry Haddow and Faheem Kirefu. 2020. Pmindia - a collection of parallel corpora of languages of India. ArXiv, abs/2001.09907.
2. Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senel-lart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In Proceedings of ACL 2017, System Demonstrations, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.
3. Haiyue Song, Raj Dabre, Zhuoyuan Mao, Fei Cheng, Sadao Kurohashi, Eiichiro Sumita. 2020. Pre-training via Leveraging Assisting Languages for Neural Machine Translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop pages 279–285. Association for Computational Linguistics.
4. Surafel Melaku Lakew, Mauro Cettolo, Marcello Federico. 2018. A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation.