A3-108 Machine Translation System for LoResMT 2019

Saumitra Yadav **Vandan Mujadia** **Manish Shrivastava**
Machine Translation - Natural Lanugae Processing Lab
Language Technologies Research Centre
Kohli Center on Intelligent Systems
International Institute of Information Technology - Hyderabad
saumitra.yadav,vandan.mu{@research.iiit.ac.in}
m.shrivastava@iiit.ac.in

Abstract

In this paper, we describe our machine translation systems submitted to LoResMT 2019 Shared Task. Systems were developed for Bhojpuri, Magahi, Sindhi, Latvian $\Longleftrightarrow$ (English). This paper outlines preprocessing , configuration of the submitted systems and the results produced using the same.

1 Introduction

The task of Machine Translation aims to obtain valid translation of text of one language to another. Data driven MT system uses parallel sentences (i.e, x^{th} sentences in two languages show same meaning). For the data driven system to learn translation, it requires sufficient amount of parallel text (bi-text) (Turchi et al., 2008), which is not always easy to get. Scarcity of parallel text can hinder data driven systems ability to give decent translations (Koehn and Knowles, 2017).

For languages like Bhojpuri, Sindhi and Maghai which are primarily spoken in northern India by around 50 million, 1.6 million, 12 million people respectively[1] resources are scarce to obtain a decent machine translation system. As for Latvian, which is spoken by roughly 1.75 million people primarily in Latvia and is one of the official languages of the EU[2]. In LoResMT 2019, we participated as team A3-108 and trained 24 systems for English to (Bhojpuri, Magahi, Sindhi, Latvian) and vice-versa with 3 systems for each direction.

[1]http://www.censusindia.gov.in/2011Census/Language-2011/Statement-1.pdf

[2]https://www.ethnologue.com/18/language/lav/

2 Data

Parallel and monolingual corpora for Bhojpuri, Magahi and Sindhi received for the shared task. Monolingual data for English and Latvian were taken from Goldhahn et al (2012). We included training data to the monolingual corpus of each language for decent language model. Statistics of parallel and monolingual text are presented in Table 1 and 2 respectively.

Language Pair	Train	Dev	Test
eng-bho	28999	500	250
eng-mag	3710	500	250
eng-sin	29014	500	250
eng-lav	54000	1000	500

Table 1: English-low resources languages (eng-English, bho-Bhojpuri, mag-Magahi, sin-Sindhi and lav-Latvian corpus) split statistics. Number indicates number of parallel sentences.

Language	# of sentences
bho	78999
mag	19027
sin	102345
lav	2053998
eng	2410767

Table 2: We concatenate training data with monolingual data for (eng-English, bho-Bhojpuri, mag-Magahi, sin-Sindhi and lav-Latvian corpus).

3 System Description

We utilize both Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) with attention for our systems. Following subsections describe steps involving preprocessing and training configurations for NMT and SMT.

3.1 Preprocessing

Following are the preprocessing steps for both SMT and NMT.

- **Tokenization**: We use IndicNLP Toolkit[3] to tokenize Bhojpuri, Maghai and Sindhi (train, dev, test and monolingual) as a first step. For English and Latvian, we utilize default Moses toolkit[4](Koehn et al., 2007) tokenizer to obtain clean tokenized text.
- Also, for English, we keep letter case as it is to capture syntactic importance e.g. *The* is at start of sentence would roughly be the determinant of subject unlike *the* in the middle of a sentence and to help translate Named entity.

3.2 Training configuration for Neural Machine Translation

NMT make use of neural networks to learn to generate most likely text sequence as output given input text sequence(Sutskever et al., 2014; Bahdanau et al., 2014). Recent work in machine translation make use of self attention(Vaswani et al., 2017) to achieve State of Art results for resource rich language pairs. Due to low resource settings (Koehn and Knowles, 2017), we avoid the use of transformer and explore sequence to sequence with attention architecture (Bahdanau et al., 2014) for our NMT based systems. We make use of Nematus toolkit[5](Sennrich et al., 2017) to carry out our NN based experiments for all 8 directions (English $\Longleftrightarrow$ Bhojpuri, English $\Longleftrightarrow$ Magahi, English $\Longleftrightarrow$ Sindhi and English $\Longleftrightarrow$ Latvian).

In Table 3, Columns show total number of unique words with minimum count (mc) 2 and 1 in training text for respective language pairs (L1-L2). One can observe that there is a significant increase in unique count between mc>=2 and mc>=1. Hence, vocabulary size increases significantly which affects learning due in low resource settings (because almost half of the vocab has frequency 1). Therefore, we explore Byte Pair Encoding (BPE) (Sennrich et al., 2015) to handle rare words effectively.

Following are hyper-parameters we use in our NMT systems and rest were default as mentioned in Nematus,

- BPE Merge Operations: 5000
- Hidden Layer Dimension of LSTM: 200
- Loss: cross entropy
- Optimizer: Adam
- Beam Size (During Training): 4
- Beam Size (During Testing): 10
- Size of Embedding Layer for **Method1-a**: 50
- Size of Embedding Layer for **Method2-a**: 200

Also, we train two systems in each direction English<>(Bhojpuri, Magahi, Sindhi, Latvian) by keeping dimension of embedding layer to 50 and 200 respectively. We use Adam Optimizer(Kingma and Ba, 2014) with cross entropy loss across all systems.

Language Pair L1 - L2	# of unique words			
	mc>=2		mc >=1	
	L1	L2	L1	L2
eng-bho	6710	8790	12684	19754
eng-mag	2946	3355	5650	6504
eng-sin	6726	7651	12127	15689
eng-lav	16145	32248	27896	60376

Table 3: Number of Unique words in training data for language pairs (eng-English, bho-Bhojpuri, mag-Magahi, sin-Sindhi and lav-Latvian), with minimum count (mc) >=2 and >=1 .

3.3 Training configuration for Statistical Machine Translation

Phrase Based Statistical Machine Translation (PB-SMT) is a statistical approach which uses co-occurrence of word sequences across parallel text to learn translation probabilities. SMT utilizes aforementioned probabilities and language model to generate translation text given an input text (Koehn et al., 2003). We make use of Moses toolkit (Koehn et al., 2007) for this paradigm. We also use GIZA++ (Och and Ney, 2003) to find alignments between parallel text and grow-diag-final-and method (Koehn et al., 2003) to extract aligned phrases. We utilize KenLM (Kenneth Heafield, 2011) to train a trigram model with kneser ney smoothing on monolingual corpus of all languages and MERT (Och, 2003) is used for tuning the trained models (named as **Method3-b** in results).

[3] http://anoopkunchukuttan.github.io/indic_nlp_library/
[4] https://github.com/mosessmt/mosesdecoder
[5] https://github.com/EdinburghNLP/nematus

Experiment	BLEU	Precision	Recall	F-Measure
Bho2Eng-Method1-a	10.12	16.27	15.46	15.85
Bho2Eng-Method2-a	12.09	18.72	17.67	18.18
Bho2Eng-Method3-b	**17.03**	**22.28**	**22.43**	**22.35**
Eng2Bho-Method1-a	6.19	12.52	11.59	12.04
Eng2Bho-Method2-a	10.5	**18.11**	15.34	16.61
Eng2Bho-Method3-b	**10.69**	16.74	**17.07**	**16.9**
Eng2Lav-Method1-a	17.06	26.74	21.05	23.56
Eng2Lav-Method2-a	28.46	33.71	32.19	32.93
Eng2Lav-Method3-b	**33.78**	**37.75**	**38.55**	**38.15**
Eng2Mag-Method1-a	1.63	8.66	5.95	7.05
Eng2Mag-Method2-a	1.83	9.13	5.09	6.54
Eng2Mag-Method3-b	**9.37**	**16.21**	**17.06**	**16.62**
Eng2Sin-Method1-a	17.43	22.2	22.91	22.55
Eng2Sin-Method2-a	25.17	30.09	29.09	29.58
Eng2Sin-Method3-b	**37.58**	**40.4**	**40.52**	**40.46**
Lav2Eng-Method1-a	31.79	38.45	35.11	36.7
Lav2Eng-Method2-a	37.27	42.68	40.42	41.52
Lav2Eng-Method3-b	**43.6**	**46.86**	**47.59**	**47.22**
Mag2Eng-Method1-a	1.86	8.58	6.37	7.31
Mag2Eng-Method2-a	3.03	10.28	6.67	8.09
Mag2Eng-Method3-b	**9.71**	**16.55**	**17.15**	**16.84**
Sin2Eng-Method1-a	19.11	25.54	24.01	24.75
Sin2Eng-Method2-a	26.68	32.38	30.81	31.58
Sin2Eng-Method3-b	**31.32**	**36.06**	**35.86**	**35.96**

Table 4: Performace of translation systems in terms of BLEU score, Precision, Recall and F-Measure

4 Result

Table 4 shows performance of 24 systems in terms of BLEU (Papineni et al., 2002) score, Precision, Recall and F-Measure. First column (*Experiment field*) shows the language direction and method used. From the table 4, we can see that for each language direction we report three different experiments(1,2 for NMT and 3 for SMT) as described in Section-3.

From the experiments, We observe that SMT is consistently outperforming NMT in low resource settings (Table 4).

- hyperparameters of network along with mention of method 1 and 2
- mention of method 3 in smt

References

Bahdanau, Dzmitry and Cho, Kyunghyun and Bengio, Yoshua 2014. *Neural machine translation by jointly learning to align and translate*, arXiv preprint arXiv:1409.0473

Goldhahn, Dirk and Eckart, Thomas and Quasthoff, Uwe. 2012. *Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages.* LREC. Volume 29. 31–43

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

Kingma, Diederik P and Ba, Jimmy 2014. *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980

Koehn, Philipp and Hoang, Hieu and Birch, Alexandra and Callison-Burch, Chris and Federico, Marcello and Bertoldi, Nicola and Cowan, Brooke and Shen, Wade and Moran, Christine and Zens, Richard and others 2007. *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 177–180

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

Koehn, Philipp and Knowles, Rebecca 2017. *Six challenges for neural machine translation*, arXiv preprint arXiv:1706.03872

Och, Franz Josef and Ney, Hermann 2003. *A Systematic Comparison of Various Statistical Alignment Models*, Computational Linguistics Number 1, Volume 29 19–51

Och, Franz Josef 2003. *Minimum error rate training in statistical machine translation*, Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1 160–167 Association for Computational Linguistics

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

Sennrich, Rico and Haddow, Barry and Birch, Alexandra 2015. *Neural machine translation of rare words with subword units*, arXiv preprint arXiv:1508.07909

Sennrich, Rico and Firat, Orhan and Cho, Kyunghyun and Birch, Alexandra and Haddow, Barry and Hitschler, Julian and Junczys-Dowmunt, Marcin and Läubli, Samuel and Miceli Barone, Antonio Valerio and Mokry, Jozef and Nadejde, Maria 2017. *Nematus: a Toolkit for Neural Machine Translation*, Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics 65–68 Association for Computational Linguistics

Sutskever, Ilya and Vinyals, Oriol and Le, Quoc V 2014. *Sequence to sequence learning with neural networks*, Advances in neural information processing systems 3104–3112

Turchi, Marco and De Bie, Tijl and Cristianini, Nello 2008. Alternation. *Learning performance of a machine translation system: a statistical and computational analysis*, Proceedings of the Third Workshop on Statistical Machine Translation 35–43 Association for Computational Linguistics

Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, Łukasz and Polosukhin, Illia 2017. *Attention is all you need*, Advances in neural information processing systems 5998–6008