A3-108 Machine Translation System for LoResMT Shared Task @MT Summit 2021 Conference

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

Abstract

In this paper, we describe our submissions for LoResMT Shared Task @MT Summit 2021 Conference. We built statistical translation systems in each direction for English $\Longleftrightarrow$ Marathi language pair. This paper outlines initial baseline experiments with various tokenization schemes to train models. Using optimal tokenization scheme we create synthetic data and further train augmented dataset to create more statistical models. Also, we reorder English to match Marathi syntax to further train another set of baseline and data augmented models using various tokenization schemes. We report configuration of the submitted systems and results produced by them.

1 Introduction

Machine Translation systems are systems which translate from source language to target. There are multiple ways of creating such a system - rule based, data driven, hybrid etc. We are using data driven methods to create translation system. In data driven methods - statistical (Koehn et al., 2003) and neural methods (Bahdanau et al., 2014) have been employed to build decent MT systems in resource setting like English $\Longleftrightarrow$ French. In LoResMT shared task (Ojha et al., 2021) we are dealing with low resource setting for English, Marathi pair. According to Koehn and Knowles (2017), compared to statistical methods neural methods have a drawback when used in low resource setting. Hence, for this shared task we are using only phrase based statistical models to build translation models using Moses[1] (Koehn et al., 2007).

Marathi is morphologically richer, agglutinative language when compared to English. Also, former follows SOV as canonical syntactic structure while latter follows SVO. Level of difference in morphological richness and syntactic divergence between the two languages suggests to look for methods which can help to address them to certain extent in phrase based statistical models. Since we are in low resource setting, to address data sparsity problem, we use various tokenization schemes, e.g. BPE (Sennrich et al., 2016b), morfessor (Virpioja et al., 2013). Combinations of these tokenization schemes are used with SMT based method to create a baseline systems. After checking the optimal tokenization scheme, we use that scheme to augment training data with synthetic dataset using back translation (Sennrich et al., 2016a). As was the case in baseline systems, augmented dataset goes through prepossessing with various tokenization schemes and SMT method to build more systems. We elevate the amount of learning, the reordering model of SMT has to do, by making use of rule based reordering system

[1] http://statmt.org/moses/

Dataset	Baseline and Reordered Baseline		Augmented and Reordered Augmented	
Language	English	Marathi	English	Marathi
Monolingual	34891	40972	56789	57569
training	20651		59146	
dev	500		500	

Table 1: Data statistics, number of sentences for each set of experiments

(Patel et al., 2013), (Kunchukuttan et al., 2014) to reorder English to match Marathi syntax. With this we build another set of baseline systems for reordered English, Marathi pair. Like in baseline systems, mentioned above, here also we make use of various tokenization schemes. After comparing these schemes, we create synthetic dataset using back translation to augment reordered English, Marathi pair. Subsequent sections give more detailed overview of the systems developed.

2 SMT Systems

We use SMT model to make initial baseline systems using various tokenization schemes. We further make use of rule based reordering model to create another set of baseline systems using reordered English, Marathi pair. These two sets of systems are then used to create synthetic data set for data augmentation to train SMT models.

2.1 Data

For this shared task organisers provided parallel and monolingual corpus. We include Marathi training, dev dataset to already existing monolingual coprus to create Marathi monolingual corpus. For English monolingual corpus we joined English training and dev data from both (English $\Longleftrightarrow$ Marathi, English $\Longleftrightarrow$ Irish) language pair provided by organizers. As a first preprocessing step, we used the IndicNLP toolkit[2] to tokenize Marathi and Moses tokenizer[3] to tokenize English. Then we learned subwords using Byte pair encoding (Sennrich et al., 2016b) with 10000 merge operations on monolingual corpus and tokenized training and dev accordingly. We also used morfessor (Virpioja et al., 2013) as an alternative tokenization scheme. Morfessor model was also trained on full monolingual corpus. Table 1 provides statistics of datasets processed.

We made use of CFILT toolkit[4] to preorder English sentence in train, dev and monolingual text. Similar to previous sets of baseline systems, we use vairous tokenization schemes - moses tokenizer, BPE, Morfessor and train another set of baseline systems. Table 1 provides you with statistics of reordered English. We used all possible combination of tokenization schemes while training all models. These tokenization schemes are named as follow,

- BasicTok: Basic Tokenization using Indic NLP for Marathi and Moses tokenizer for English.
- BPE: text tokenized using BPE into subword.
- Morf: text tokenized using morfessor.

[2] https://anoopkunchukuttan.github.io/indic_nlp_library/

[3] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

[4] https://www.cfilt.iitb.ac.in/static/download.html

2.2 Translation Models

We made use of Moses toolkit (Koehn et al., 2007) to build statistical models trained with various tokenized bitext pairs. 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. And utilize KenLM (Heafield, 2011) to train a trigram model with kneser ney smoothing on monolingual corpus of both languages. MERT (Och, 2003) is used for tuning the trained models. We also trained a reordering system for Reordered English to English so that we can have Reordered English as pseudo-pivot language.

2.2.1 Transliteration Module

Since we are building systems in low resource setting, its entirely possible to get unknown words while translating. To see if we can also counter unknown words in this resource constrained environments, we also made a small transliteration system. First a phrase based model was trained on English Marathi bitext using Moses(Koehn et al., 2007) with max phrase length[5] set to 1 to find tokens with very high alignment probability (we took average of 4 probabilities and took token pair with value > 0.79). We got 1557 pairs of tokens, we tokenized them character wise and used 1500 for training and 57 for tunning to build a transliteration system (by posing transliteration as translation problem). Since transliteration system is trained on very small corpus and hence prone to error, for each output from SMT translation system we give two outputs. One in which we made use of transliteration system for unknown words and another one in which we did not.

2.2.2 Performance on Dev sets

We used dev set to evaluate above mentioned models and all models which are described later on. Outputs were post processed according to the tokenization scheme of respective target language in each model, and then detokenized. After evaluating all systems using sacrebleu (Post, 2018), Table 2 lists the result on dev sets for baseline systems trained in English to Marathi Direction and Table 3 lists the result of systems trained on Marathi to English direction. If we

Tokenization Scheme	Baseline SMT		Augment SMT		Baseline Reordered SMT		Augment Reordered SMT	
	unk	transli-terate	unk	transli-terate	unk	transli-terate	unk	transli-terate
EnTok MrTok	58.2	58.1	57.6	57.5	59.2	59.1	62.7	62.7
EnBPE MrBPE	50.1	50.1	52.0	51.9	53.1	53.1	56.7	56.7
EnMorf MrMorf	41.3	41.3	43.6	43.6	45.7	45.6	51.3	51.3
EnTok MrBPE	49.3	49.1	50.7	50.6	51.5	51.4	54.7	54.7
EnTok MrMorf	47.4	47.3	47.3	47.2	49.8	49.7	54.7	54.7
EnBPE MrTok	54.3	54.3	54.4	54.4	56.7	56.7	58.9	58.9
EnBPE MrMorf	46.0	46.0	48.0	48.0	49.5	49.5	53.2	53.2
EnMorf MrTok	51.4	51.3	50.9	50.9	53.7	53.6	55.6	55.6
EnMorf MrBPE	44.3	44.2	45.9	45.8	47.8	47.8	52.3	52.3

Table 2: Results of systems for English To Marathi language direction. unk column contain output of systems where unknown were kept as they are, in transliterate column they were transliterated using small transliteration system

look at table 2, we can see that using reordering as preprocessing tool was helpful to system

[5] http://www.statmt.org/moses/?n=FactoredTraining.TrainingParameters

Tokenization Scheme	Baseline SMT		Augment SMT		Baseline Reordered SMT		Augment Reordered SMT	
	unk	transli-terate	unk	transli-terate	unk	transli-terate	unk	transli-terate
EnTok MrTok	70.4	70.4	72.4	72.4	55.6	55.7	61.3	61.3
EnBPE MrBPE	62.6	62.6	64.5	64.5	55.6	55.6	56.6	56.6
EnMorf MrMorf	56.0	56.0	57.5	57.5	51.1	51.1	53.4	53.4
EnTok MrBPE	62.0	62.0	64.8	64.8	54.4	54.4	56.0	56.0
EnTok MrMorf	62.9	62.9	63.6	63.6	55.5	55.5	57.5	57.5
EnBPE MrTok	67.9	68.0	69.6	69.6	55.6	55.5	58.1	58.1
EnBPE MrMorf	61.5	61.5	62.7	62.7	54.4	54.4	57.6	57.6
EnMorf MrTok	61.3	61.4	62.9	62.9	51.7	51.7	54.1	54.1
EnMorf MrBPE	54.9	54.9	59.1	59.1	51.3	51.3	52.3	52.3

Table 3: Results of systems for Marathi To English language direction. unk column contain output of systems where unknown were kept as they are, in transliterate column they were transliterated using small transliteration system

translating in English to Marathi direction. Whereas, training on Marathi to reordered English (Table 3) didnt get same positive result. Also surprising was dip in BLEU scores when using subwords. Using baseline systems with BasicTok as tokenization scheme for both scenarios (in both English and reordered English scenario) we created synthetic datasets using backtranslation(Sennrich et al., 2016a). Statistics for augmented datasets are given in Table 1. We used augmented data set with Moses to build SMT systems. Moses was used in same configuration as before. We employed all tokenization schemes combinations and result of same on dev sets are available in Table 2 and 3. Similar to trend seen in baseline systems on dev datasets, here also Augmented Reordered English to Marathi produce better score that Augmented English to Marathi. Marathi to English was better than Marathi to Reordered English to English. In most of the systems transliteration module was not helpful.

3 Result

For each language direction we submitted 72 output files. Table 4 shows the scores of top 3 systems for each direction. In case of English to Marathi translation direction, similar to trend seen on devsets, reordered English to Marathi systems fared better than canonical English to Marathi systems. Though tokenization scheme used was BPE for best system. While in case of Marathi to English translation direction, making a Marathi to reordered English did not preform better than Marathi to canonical English. Also we saw baseline system with BPE tokenized English and Marathi with morfessor as prepossessing step was better than all other system configurations, followed by Augmented Marathi with BPE to English with BPE . In terms of comparison to other teams, although our Marathi to English systems did not fare well, we were in top 3 for English to Marathi systems under constrained conditions.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.

Description of Translation System	Tokenization Scheme		Scores		
	Source	Target	BLEU	TER	CHRF
Augmented Reorder English to Marathi SMT system	BPE	BPE	11.8	0.45	0.95
Augmented Reorder English to Marathi SMT system	BPE	BPE	11.8	0.45	0.95
Baseline Reordered English to Marathi SMT system	basicTok	basicTok	11.4	0.43	0.934
Baseline Marathi to English SMT System	Morf	BPE	14.6	0.47	0.945
Baseline Marathi to English SMT System	Morf	BPE	14.6	0.47	0.945
Augmented Marathi to English SMT System	BPE	BPE	14.5	0.42	0.866

Table 4: Result of our top 3 systems on testsets in each translation direction

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST.

Kunchukuttan, A., Mishra, A., Chatterjee, R., Shah, R., and Bhattacharyya, P. (2014). Sata-anuvadak: Tackling multiway translation of indian languages. *pan*, 841(54,570):4–135.

Och, F. J. (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.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Ojha, A. K., Liu, C.-H., Kann, K., Ortega, J., Satam, S., and Fransen, T. (2021). Findings of the LoResMT 2021 Shared Task on COVID and Sign Language for Low-Resource Languages. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages*.

Patel, R. N., Gupta, R., Pimpale, P. B., and M, S. (2013). Reordering rules for English-Hindi SMT. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, pages 34–41, Sofia, Bulgaria. Association for Computational Linguistics.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Virpioja, S., Smit, P., Grönroos, S.-A., Kurimo, M., et al. (2013). Morfessor 2.0: Python implementation and extensions for morfessor baseline.