JHU LoResMT 2019 Shared Task System Description

Paul McNamee
Johns Hopkins University
Human Language Technology Center of Excellence
810 Wyman Park Dr., Baltimore, Maryland 21211 USA
mcnamee@jhu.edu

Abstract

We describe the JHU submission to the LoResMT 2019 shared task, which involved translating between Bhojpuri, Latvian, Magahi, and Sindhi, to and from English. JHU submitted runs for all eight language pairs. Baseline runs using phrase-based statistical machine translation (SMT) and neural machine translation (NMT) were produced. We also submitted neural runs that made use of backtranslation and ensembling. Preliminary results suggest that system performance is reasonable given the limited amount of training data.

1 Introduction

JHU submitted runs for each of the eight language pairs in the shared task. A goal of our participation was to compare baseline SMT and NMT systems in low resource conditions. For the most part we used homogenous processing for our runs involving different language pairs. However, our primary interest was exploring translation to English, and we paid more attention and submitted more runs for those conditions. Also, there was so little data for Magahi, that using different hyperparameters seemed well-motivated. We used monolingual English data in some of our submissions, but did not make use of the monolingual data provided in other languages. Our team code was L19T5.

2 Data

The amount of provided parallel data, by language, is shown in Table 1. Note, the provided Sindhi data was marked as "sin", however the ISO-639-3 code for Sindhi is "snd". We use "sin" throughout for consistency with the shared task.

Pair	Train	Tune	Test
bho–eng	28,999	500	250
lav–eng	54,000	1,000	500
mag–eng	3,710	500	250
sin–eng	29,014	500	250

Table 1: Number of parallel sentences used for each language pair, by partition. Test sets with English as the source language had the same size, except for eng–sin which had a test set of 249 sentences.

3 Models

In this section we describe the methods used to produce submissions to the task. Where English was the source language we used a SMT baseline to produce one submission, and we used NMT to both produce a submission and to translate 100,000 English sentences to the source language for subsequent use in backtranslation experiments. Characteristics of the submissions are shown in Table 2 and Table 3.

3.1 SMT Baseline

A phrase-based SMT system, Apache Joshua (Post et al., 2015), was used for Condition A[1] and for Condition C[2]. Sentences were tokenized using the Moses tokenizer and lower-cased (when appropriate). Sentences longer than 75 tokens in length were ignored during training. KenLM (Heafield, 2011) was used to train 4–gram language models using the target side of training bitext. When translating to English, a larger language model based on

© 2019 The authors. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

[1] Provided corpora only
[2] Use of publicly available corpora

Run	Cond	Type	Aux. LM	BPE units	Chkpt freq
L19T5-bho2eng-pbmt-a	A	SMT	–	–	–
L19T5-bho2eng-pbmtlm-a	C	SMT	Yes	–	–
L19T5-bho2eng-xform-a	A	NMT	–	10,000	4,000
L19T5-bho2eng-xformbt-a	C	NMT	–	15,000	4,000
L19T5-lav2eng-pbmt-a	A	SMT	–	–	–
L19T5-lav2eng-pbmtlm-a	C	SMT	Yes	–	–
L19T5-lav2eng-xform-a	A	NMT	–	10,000	4,000
L19T5-lav2eng-xformbt-a	C	NMT	–	15,000	4,000
L19T5-mag2eng-pbmt-a	A	SMT	–	–	–
L19T5-mag2eng-pbmtlm-a	C	SMT	Yes	–	–
L19T5-mag2eng-xform-a	A	NMT	–	2,500	2,000
L19T5-mag2eng-xformbt-a	C	NMT	–	15,000	4,000
L19T5-sin2eng-pbmt-a	A	SMT	–	–	–
L19T5-sin2eng-pbmtlm-a	C	SMT	Yes	–	–
L19T5-sin2eng-xform-a	A	NMT	–	10,000	4,000
L19T5-sin2eng-xformbt-a	C	NMT	–	15,000	4,000

Table 2: Characteristics of submitted runs with English as the target language. Note, the runs labelled "xformbt-a" were named in error — they were in fact Condition C runs.

Run	Cond	Type	Aux. LM	BPE units	Chkpt freq
L19T5-eng2bho-pbmt-a	A	SMT	–	–	–
L19T5-eng2bho-xform-a	A	NMT	–	10,000	4,000
L19T5-eng2lav-pbmt-a	A	SMT	–	–	–
L19T5-eng2lav-xform-a	A	NMT	–	10,000	4,000
L19T5-eng2mag-pbmt-a	A	SMT	–	–	–
L19T5-eng2mag-xform-a	A	NMT	–	2,500	2,000
L19T5-eng2sin-pbmt-a	A	SMT	–	–	–
L19T5-eng2sin-xform-a	A	NMT	–	10,000	4,000

Table 3: Characteristics of submitted runs with English as the source language.

a 5% sample of English Gigaword 5th edition[3] was also used (10.5 million sentences, 229 million tokens).

In Condition C, no additional bitext was utilized in any of the language pairs, however, a larger target-side language model was used for models translating to English.

3.2 NMT Baseline

The second system we employed was Sockeye (Hieber et al., 2017), a sequence-to-sequence transduction model based on the Apache MXNet library. Sockeye supports CNNs, RNNs, and Transformer models. For the LoResMT shared task we used transformer models (Vaswani et al., 2017). The models used 4 stacked layers in the encoder and decoder, an embedding and model size of 512, a feed-forward hidden layer size of 1024 units, and 8 self-attention heads. Training was done with a batch size of 4,096 words, a checkpoint frequency of either 2,000 or 4,000, and an initial learning rate of 0.0002. The optimizer was Adam. Training continued until validation perplexity failed to improve for 10 consecutive checkpoints, or until the maximal number of epochs (100) was reached. Initial models were trained for Condition A in both translation directions for all four low resource languages. Text was tokenized by the Moses tokenizer, lowercased, and then BPE was applied using 2,500 to 15,000 BPE units (Sennrich et al., 2016), depending on the language and condition.

The four NMT runs for the English-to-X pairs were based on training a single model in each language. However, four independently trained models with different random initializations were used to create ensemble decodes in the X-to-English pairs. Sockeye provides support for ensemble decoding by combining output layer probabilities from separate training instances.

3.3 NMT with Backtranslation

In Condition C we again used no additional bitext, however, these neural runs used 100,000 sentences randomly drawn from our English Gigaword subsample to create synthetic bitext using backtranslation with an English-to-X model used for Condition A. These machine-produced translations were then used with the provided bitext to build X-to-English models, and inference was again performed using an ensemble of four separate models. Our interest was in seeing whether backtranslation would provide gains in very low resource settings.

[3]LDC2011T07

4 Results and Discussion

All of our runs with English as the source language were Condition A (*i.e.,* provided data only). Results for these runs are shown in Table 4. We observe that phrase-based MT outperformed neural MT in all four low-resource scenarios, which is not too surprising given the limited amount of provided training data (refer to Table 1).

Pair	SMT	NMT
eng–bho	**3.01**	1.00
eng–lav	**23.24**	13.22
eng–mag	**5.66**	1.74
eng–sin	**7.72**	3.08

Table 4: Baseline SMT (pbmt) and NMT (xform) runs where English was the source language. All runs are Condition A.

Results with English as the target language are shown in Table 5.

Pair	SMT	SMT+LM	NMT	NMT+BT
bho–eng	14.20	0.14	**15.19**	13.05
lav–eng	**36.93**	1.24	34.54	35.48
mag–eng	**5.64**	0.32	4.32	1.37
sin–eng	24.55	0.11	**28.85**	23.10

Table 5: Runs for four conditions when English was the target language: SMT Baseline (A), SMT w/ auxiliary LM (C), NMT Baseline (A), and NMT using backtranslation (C).

With English as the target language, the results are mixed. SMT outperforms in two of four languages, and NMT is better in the other two. The SMT runs that used an auxiliary language model failed utterly — the results appear so poor, that it seems possible that an error was made during processing.

We observe notably higher scores in Latvian, which makes sense as it is the language pair with the greatest amount of training bitext (54,000 sentences). However, Sindhi and Bhojpuri have training sets of comparable size, yet Sindhi has appreciably higher scores.

Our recipe for backtranslation failed in three of four cases. Only in the highest resource language (*i.e.,* Latvian) did we find higher BLEU scores in our NMT models when backtranslating English text.

5 Conclusion

We created baseline SMT and NMT systems for the LoResMT 2019 shared task, and our submitted runs appeared to perform relatively well based on the preliminary results released by the task organizers. While language model augmentation failed to improve SMT performance for as yet undetermined reasons, use of backtranslation was successful in the highest resource language setting. In general, the statistical models outperformed the neural models in these low resource settings, a finding consistent with other reports in the literature (Koehn and Knowles, 2017).

References

Heafield, Kenneth. 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.

Hieber, Felix, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *CoRR*, abs/1712.05690.

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

Post, Matt, Yuan Cao, and Gaurav Kumar. 2015. Joshua 6: A phrase-based and hierarchical statistical machine translation system. *The Prague Bulletin of Mathematical Linguistics*, 104(1):5–16.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. 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, August. Association for Computational Linguistics.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Guyon, I., U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.