Factored Neural Machine Translation at LoResMT 2019

Saptarashmi Bandyopadhyay
Department of Computer Science and Engineering
Pennsylvania State University, University Park
University Park, P.A., U.S.A.
sbandyo20@gmail.com

Abstract

Low resource languages face a major challenge in developing machine translation systems due to unavailability of accurate and parallel datasets with a large corpus size. In the present work, Factored Neural machine Translation Systems have been developed for the following bidirectional language pairs: English & Bhojpuri, English & Magahi, English & Sindhi along with the uni-directional language pair English - Latvian. Both the lemma and Part of Speech (PoS) tags are included as factors to the surface-level English words. No factoring has been done on the low resource language side. The submitted systems have been developed with the parallel datasets provided and no additional parallel or monolingual data have been included. All the seven systems have been evaluated by the LoResMT 2019 organizers in terms of BLEU score, Precision, Recall and F-measure evaluation metrics. It is observed that better evaluation scores have been obtained in those MT systems in which English is the target language. The reason behind this is that the incorporation of lemma and pos tags factors for English words has improved the vocabulary coverage and has also helped in generalization. It is expected that incorporation of linguistic factors on the low resource language words would have improved the evaluation scores of the MT systems involving those languages on the target side.

1 Introduction

Data driven machine translation systems do not perform well involving Low Resource (LowRes languages since less parallel data are publicly available for these languages. However, limited monolingual data along with language analysis tools with acceptable performance measures are available for such languages. Incidentally, a large number of people use such low resource languages.

Neural machine translation (NMT) systems are the current state-of-the-art systems as the translation accuracy of such systems is very high for languages with large amount of training corpora being available publicly. Current NMT Systems that deal with LowRes languages (Guzman et. al., 2019; AMTA, 2018) are based on unsupervised neural machine translation, semi-supervised neural machine translation, pretraining methods leveraging monolingual data and multilingual neural machine translation among others.

Meanwhile, research work on Factored NMT systems (Koehn and Knowles, 2017; Garcıa-Martınez et. al. 2016; Senrich and Haddow, 2016) have evolved over the years. The factored NMT architecture has played a significant role in increasing the vocabulary coverage over standard NMT systems. The syntactic and semantic information from the language is useful to generalize the neural models being learnt from the parallel corpora. The number of unknown words also decreases in Factored NMT systems.

In the present work, the idea of using factored neural machine translation has been explored in the 7 machine translation systems. The parallel corpus has been augmented to include factors like Lemma (using Porter Stemmer) and PoS tags (using TnT Tagger) for English words. No factoring has been done on the low resource language side. After factoring is done, the training dataset has been tokenized and byte pair encoding has been implemented, thereafter.

2   Related Works

The major research areas in low resource MT systems are described in (Guzman et. al., 2019; AMTA, 2018). One major area of research is to effectively use available monolingual data. It includes semi-supervised methods relying on backtranslation, integration of a language model into the decoder extending to unsupervised approaches that use monolingual data both for learning good language models and for creating artificial parallel data.

Another primary area of research is to work on a weakly supervised learning setup in which original parallel training corpus is augmented with comparable corpora. NMT systems for low resource languages have been developed in (Guzman et. al., 2019; AMTA, 2018) in four learning settings, semi-supervised in which monolingual data is utilized on the target side, weakly supervised setting in which noisy comparable corpora is used, fully unsupervised setting in which only monolingual data on both the source and the target sides are used to train the model and the supervised model in which only parallel corpus is used during training.

The vocabulary coverage increases significantly in factored neural architecture (Koehn and Knowles, 2017; Garcıa-Martınez et. al.. 2016; Senrich and Haddow, 2016) while decreasing the number of unknown words. The linguistic decomposition of the words in terms of factors like lemma, PoS tags and other grammatical information can be applied on the source or on the target side or on both the sides.

According to the literature survey factored NMT system has not yet been applied to MT system development in Low Resource languages.

3   Factored Neural Machine Translation System

The following language pairs have been considered for development of Factored Neural Machine Translation systems:
1) English to Bhojpuri
2) Bhojpuri to English
3) English to Magahi
4) Magahi to English
5) English to Sindhi
6) Sindhi to English
7) English to Latvian

Only the provided corpora has been used for translation in all cases. The English side of the parallel corpora has been factored with the lemma and Part-of-Speech(PoS) tag of the surface word in all the 7 language pairs. The English lemma has been obtained using the Porter Stemmer (Porter, 1980). The TnT tagger has been used to obtain the PoS tags of English words (Brants, 2000). An example is as follows: the factor information of the surface word 'When' is obtained and augmented as 'When|when|WRB', where 'when' is the lemma and 'WRB' or 'Wh-adverb' is the PoS tag for the surface-level English word 'When'. No factoring has been done for the low resource language (Bhojpuri, Magahi, Sindhi, Latvian) side of the parallel corpora. Then the model for byte pair encoding (BPE) is trained with the training corpus on the source and target sides for all the language pairs. The vocabulary for byte pair encoding (BPE) is constructed with 32000 vocabulary size. Pre-tokenization has not been done as sentencepiece[1] tool has been used which does not always require pre-tokenization. The source and the target sides of the parallel corpora are then encoded using the model constructed by sentencepiece[1]. These datasets are used for training the neural model for translation. The parameters for training the neural model for translation for each of the language pairs are:
i)   Drop-out rate = 0.3
ii) 2 layered unidirectional recurrent neural network with Long Short Term Memory (LSTM) as the recurrent unit
iii) Batch size = 128 and 500 hidden units
iv) 14000 training steps
v)   Beam search as inference mode with a beam width of 5 and a length penalty weight and a coverage penalty weight of 0.4 each.

After the model is trained, the test dataset on the source side of the language pair is used to obtain the output dataset on the target side of the language pair. Once testing is done, the data is again decoded by sentencepiece[1] using the trained BPE model before. Thus, Method1 is achieved for language pairs where the low resource language is on the target side. When English is on the target side of the language pair, the generated dataset is subjected to post-processing to remove the factored information of lemma and PoS tag in it. This is referred to as Method1 for language pairs where English is on the target side. Method2 is a slight modification of Method1 where the space before punctuations ('.',',',';',':',''',''' and '!') are removed in case of language pairs where English is on the target side. For 3 low resource languages,

---

Bhojpuri, Magahi, and Sindhi, the spaced before certain punctuation marks ('|' and '!') are removed in order to study the impact of the punctuations on the BLEU scores. This is called Method2 for language pairs where the low-resource languages Bhojpuri, Sindhi and Magahi are on the target side.

4   System Evaluation Results

The results for the 7 language pairs have been illustrated in this section. It has been observed that Method 1 and Method 2 are leading to the same BLEU score, precision, recall and F-measure scores. It implies that the removal of the space character before certain punctuation marks do not have any effect on the Bleu score. Hence, the method column in the subsequent result tables have not been mentioned. The result of the Best Team for the specific language pair has been included. Since, no details are available about the specific method used by the Best team, no direct comparison has been made.

Team	BLEU score	Precision	Recall	F-measure
My Team (L19T6)	6.83	11.73	11.59	11.6
Best Team (L19T2)	10.69	16.74	17.07	16.9

Table 1: English-Bhojpuri FNMT System Results

The BLEU score for English-Bhojpuri language pair has been the second best among all the submissions. The Bleu score of the submitted system is 36% below the Best Team Score.

The Bhojpuri to English language pair also exhibits a good performance in the BLEU score. It is observed that higher Bleu scores are obtained with English as the target language. The Bleu score of the submitted system is 21% less than the Best Team Score. The precision score is only 6.5% less than that of the Best Team.

Team	BLEU score	Precision	Recall	F-measure
My Team (L19T6)	13.39	20.84	17.41	18.99
Best Team (L19T2)	17.03	22.28	22.43	22.35

Table 2: Bhojpuri-English FNMT System Results

The Bleu score for English-Sindhi submitted system is 59% lower than the Best Team System score, as shown in Table 3.

Team	BLEU score	Precision	Recall	F-measure
My Team (L19T6)	15.34	21.02	20.26	20.63
Best Team (L19T2)	37.58	40.4	40.52	40.46

Table 3: English-Sindhi FNMT System Results

Team	BLEU score	Precision	Recall	F-measure
My Team (L19T6)	26.2	33.24	29.54	31.28
Best Team (L19T2)	31.32	36.06	35.86	35.96

Table 4: Sindhi-English FNMT System Results

The Sindhi to English language pair also exhibits a good performance in the BLEU score. It is observed that higher Bleu scores are obtained with English as the target language. The Bleu score of L19T6 is 16% less than the Best Team Score.

Team	BLEU score	Precision	Recall	F-measure
My Team (L19T6)	9.02	12.01	15.43	13.41
Best Team (L19T1)	48.88	51.09	51.19	51.14

Table 5: English-Latvian FNMT System Results

The Bleu score for English-Latvian submitted system is 82% lower than the Best Team System score. It demonstrates that simply using the parallel corpus in the MT system does not always provide better result.

Team	BLEU score	Precision	Recall	F-measure
My Team (L19T6)	0.24	5.82	3.48	4.36
Best Team (L19T2)	9.37	16.21	17.06	16.62

Table 6: English-Magahi FNMT System Results

The performance of the English-Magahi is worse as the Bleu score of the submitted system is 97% below the Best Team score. However, the F-measure of the submitted system is 74% below the Best Team score. Thus, there is a better correlation with human judgment.

Team	BLEU score	Precision	Recall	F-measure
My Team (L19T6)	0.13	3.91	2.5	3.05
Best Team (L19T2)	9.71	16.55	17.15	16.84

Table 7: Magahi-English FNMT System Results

The performance of the Magahi - English is similarly worse as the Bleu score of the submitted system is 98% below the Best Team score. However, the F-measure of the submitted system is 82% below the Best Team score. Thus the correlation with human judgment is comparatively higher.

5 Conclusion

Factored Neural Machine Translation systems have been developed for the following Bidirectional language pairs: English & Bhojpuri, English & Sindhi, English & Magahi and English-Latvian. All the languages except English are Low Resource languages in which accurate and parallel datasets with larger corpus size are not available. Both the lemma and POS tags are included as factors on the English words while no factoring has been done on the low resource language side. The submitted systems have been developed only with the parallel corpus provided. Analysis of the system evaluation results demonstrate that inclusion of the lemma and PoS tags as factors on the English target side improves the Bleu score than when English is on the source side. The translation quality for English-Bhojpuri and Bhojpuri-English language pairs is very good, without using any additional dataset and by using a standard neural architecture of a 2 layered un-directional recurrent neural network, to learn the language model for translation. The lower values of the Bleu scores for the submitted systems English-Latvian, English - Magahi and Magahi-English demonstrate that using the parallel corpus only in developing the FNMT system does not improve the system evaluation scores.

References

Francisco Guzman et. al.. 2019. Two New Evaluation Data-Sets for Low-Resource Machine Translation: Nepali–English and Sinhala–English. arXiv:1902.01382v1 [cs.CL].

AMTA. 2018. Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018). *The 13th Conference of the Association for Machine Translation in the Americas.*

Phillip Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation, In *Proceedings of the First Workshop on Neural Machine Translation*, Vancouver, Canada, August 4, 28-39.

Mercedes García-Martínez et. al.. 2016. Factored Neural Machine Translation Architectures, In *Proceedings of IWSLT 2016.*

Rico Sennrich and Barry Haddow. 2016. Linguistic Input Features Improve Neural Machine Translation, In *Proceedings of the First Conference on Machine Translation*, Volume 1: Research Papers, Berlin, Germany, August 11-12, 83-91.

Thorsten Brants. 2000. TnT - A Statistical Part-of-Speech Tagger, In Proceedings of the sixth conference on Applied natural language processing, 224-231.

Porter, M. 1980. An algorithm for suffix stripping. *Program* (14.3): 130-137.