The UCF Systems for the LoResMT 2021 Machine Translation Shared Task

William Chen wchen6255@knights.ucf.edu
Brett Fazio brettfazio@knights.ucf.edu
University of Central Florida

Abstract

We present the University of Central Florida systems for the LoResMT 2021 Shared Task, participating in the English-Irish and English-Marathi translation pairs. We focused our efforts on constrained track of the task, using transfer learning and subword segmentation to enhance our models given small amounts of training data. Our models achieved the highest BLEU scores on the fully constrained tracks of English-Irish, Irish-English, and Marathi-English with scores of 13.5, 21.3, and 17.9 respectively.

1 Introduction

In this paper, we describe the systems developed at the University of Central Florida for our participation in the Marathi and Irish tasks of the LoResMT 2021 Shared Task for low-resource supervised machine translation of COVID-19 related texts (Ojha et al., 2021). For these tasks, participants were asked to develop systems for the English to Irish, Irish to English, English to Marathi, or Marathi to English translation directions. Submissions are split into three tracks based on data constraints: a constrained track using only provided data, an unconstrained track allowing for publicly available corpora, and an unconstrained track allowing for both publicly available corpora and pre-trained models.

We utilize the Neural Machine Translation (NMT) approach, due to its prevalence in current research. While they are able to achieve state-of-the-art results in high-resource translation tasks, NMT systems tend to particularly struggle in low-resource scenarios. To alleviate this, our experiments focus primarily on data augmentation and transfer learning. We tried different techniques such as back-translation and subword segmentation, although they yielded little to no improvement in most cases. Our best performing systems during development for the Irish task utilized transfer learning from English-Marathi models. For the Marathi task, the best performances came from models trained on text pre-processed with subword segmentation via a unigram language model (Kudo, 2018). We submitted six systems for evaluation, one for each translation direction in the constrained track and two unconstrained Marathi models, which achieved the highest BLEU scores in the constrained tracks of English-Irish, Irish-English, and Marathi-English.

2 Data

We only utilize the data provided by the shared task organizers (Ojha et al., 2021) for our experiments. All text was pre-processed using Moses Tokenizer (Koehn et al., 2007) and lowercased prior to training. We also experimented with a pre-trained SentencePiece (Kudo and

Richardson, 2018) tokenization model for Marathi from iNLTK (Arora, 2020). Final translation results were true-cased with Moses prior to submission.

We filtered out training sentences in the English-Marathi data that were also found in the development set from the training set to better evaluate our models during development. The parallel English-Marathi data was much larger than the English-Irish set, at 20,470 and 8,112 sentences respectively. A summary of the basic statistics of the final dataset used can be found in Table 1.

Data	Sentences	Vocabulary
EN Train (EN-GA)	8,112	15,761
GA Train (EN-GA)	8,112	17,026
EN Train (EN-MR)	20,470	27,717
MR Train (EN-MR)	20,470	42,116
EN Monolingual	8,826	20,037
MR Monolingual	21,902	39,942
EN Dev. (EN-GA)	502	2,128
GA Dev. (EN-GA)	502	2,455
EN Dev. (EN-MR)	500	3,740
MR Dev. (EN-MR)	500	4,767
EN Test (EN → GA)	500	1,912
GA Test (GA → EN)	250	1,056
EN Test (EN → MR)	500	2,344
MR Test (MR → EN)	500	2,528

Table 1: Statistics for the data used

3 System Description

We implement our models using OpenNMT-py (Klein et al., 2017). We initially tested both the Transformer (Vaswani et al., 2017) and LSTM (Hochreiter and Schmidhuber, 1997) architectures to obtain baseline results, but we found LSTMs to consistently yield better results despite attempts in optimizing Transformer parameters. As such, all of our experiments utilize LSTMs in a standard encoder-decoder setup.

We keep the majority of the parameters as their default values: we use 2 LSTM layers with 500 hidden units, an initial learning rate of 1, a dropout rate of 0.3, and stochastic gradient descent as the optimizer. We found the default step count of 100,000 to work well with the provided corpora with a batch size of 32.

3.1 Subword Segmentation

Subword segmentation is a common technique used for better dataset representation by reducing the amount of unique tokens and thus decreasing the chances of encountering unknown words. We explore Byte-Pair Encoding (BPE) (Sennrich et al., 2015) and Unigram (Kudo, 2018) segmentation, unsupervised algorithms commonly used in machine translation tasks. BPE is a greedy algorithm that initially represents a corpus at a character level, before conducting a certain number of merge operations to create subwords. Unigram initializes a large vocbulary from the corpus, before trimming down it down to meet a desired threshold.

3.2 Back-Translation

The addition of synthetic data via back-translation (Sennrich et al., 2016) has been shown to increase translation quality. To generate back-translated data, we train basic LSTM translation models on the provided parallel data. The models are then used to translate the provided monolingual data to create a parallel dataset. We also take advantage of the presence of English in both language pairs, translating the English portion of the English-Marathi training data to Irish and the English portion of the English-Irish training data to Marathi.

3.3 Transfer Learning

Transfer learning is a technique frequently used in low-resource translation, and is done by transferring the learned parameters of a high-resource parent model to low-resource child model (Zoph et al., 2016). We utilize transfer learning by training models on one language pair before fine-tuning them on the other (i.e. pre-training on English to Marathi and fine-tuning on English to Irish), leading to a total of four models trained with transfer learning: English-Marathi transferred to English-Irish, English-Irish transferred to English-Marathi, Marathi-English transferred to Irish-English, and Irish-English transferred to Marathi-English. We initialize these models with the weights of the trained LSTM baselines and fine-tune them for 100,000 steps on a new language pair. The optimizer is reset prior to fine-tuning to offset learning rate decay.

4 Experiments and Results

We first trained models using each technique to establish the effectiveness of a technique in each translation direction. Techniques that obtained a higher score than the baseline were then jointly used to develop additional models. We evaluated each model using the sacreBLEU (Post, 2018) implementation of BLEU (Papineni et al., 2002).

Model	EN→GA	GA→EN	EN→MR	MR→EN
1. LSTM	9.17	11.70	29.10	43.49
2. LSTM + BPE	8.04	10.47	27.70	39.94
3. LSTM + Unigram	8.87	9.30	28.39	43.79
4. LSTM + Back-Translation	8.75	11.32	22.26	40.29
5. LSTM + Transfer Learning	10.75	13.80	21.15	37.88
6. Model 1 + Pre-Trained Tokenizer			51.80	43.50
7. Model 2 + Pre-Trained Tokenizer			51.72	43.15
8. Model 3 + Pre-Trained Tokenizer			52.95	43.79
9. Model 4 + Pre-Trained Tokenizer			50.62	40.12
10. Model 5 + Pre-Trained Tokenizer			49.40	39.74

Table 2: BLEU scores on the validation set.

BLEU scores in the development stage are presented in Table 2. An unexpected outcome was the relative lack of benefit from subword segmentation and back-translation. We liken the former to the large vocabulary overlap between the training and validation set due to the COVID-19 specific context, as there would be fewer to no rare words that could be broken down into meaningful subwords by the segmentation algorithms.

Transfer learning from the higher resource English-Marathi models to the lower resource English-Irish models resulted in significant improvements in BLEU score (1.58 for English to Irish and 2.1 for Irish to English). However the reverse was not true, as English-Marathi models actually showed a large decrease in performance when knowledge was transferred from English-Irish.

The very high BLEU scores for the English-Marathi models can be explained by the domain overlap between data splits. We found the amount of common vocabulary between the training, development, and test sets of both language pairs to be rather large. For the Marathi texts, 94% of the development and 88% of the test vocabulary were found in the training portion.

The overlap was even more noticeable for English, with 93% of the vocabulary in development shared with training and 80% of vocabulary in testing shared with training.

Pair	Track	Model	BLEU	CHRF	TER
EN→GA	(A)	5	13.5	0.37	0.756
GA→EN	(A)	5	21.3	0.45	0.711
EN→MR	(A)	3	5.1	0.22	0.872
EN→MR	(B)	8	4.8	0.29	1.063
MR→EN	(A)	3	17.9	0.40	0.744
MR→EN	(B)	8	7.7	0.24	0.833

Table 3: Test scores via different metrics, provided by the organizers (Ojha et al., 2021)

For our final submissions, we participated in two tracks: the fully constrained track (A) and an unconstrained track (B). Track A was limited to using only data provided by the organizers. Track B allowed additional monolingual data and pre-trained models. We submitted translations generated from the systems with the highest BLEU scores in the development stage (Table 2) for each translation direction. We used LSTMs with transfer learning (Model 5) for the English to Irish and Irish to English directions, only submitting to track A. For track A of English to Marathi and Marathi to English, we used the LSTMs trained on text segmented with a unigram model (Model 3). For track B, we also used LSTMs trained on text segmented with a unigram model, but with the text pre-processed with a pre-trained language model tokenizer (Model 8). Table 3 shows the final scores of our systems on the test set, evaluated by BLEU (Papineni et al., 2002), CHRF (Popović, 2015), and TER (Snover et al., 2006).

5 Conclusion

We present systems for machine translation of Irish and Marathi to and from English. We improved over a developed baseline by incorporating transfer learning between language tasks and subword segmentation into our models. We also experimented with synthetic data generation via back-translation, which did not show any notable improvements during development. At test time, our models achieved the highest BLEU scores in the constrained tracks of English-Irish, Irish-English, and Marathi-English.

References

Arora, G. (2020). iNLTK: Natural language toolkit for indic languages. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 66–71, Online. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (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.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

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*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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. (2015). Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv:1706.03762. Version 5.*

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.