Transfer Learning across Transformer Models on Multi Language Translation Task

Kevin Ngo, Prakhar Maini

W266: Natural Language Processing
UC Berkeley School of Information
{kngo, prakharmaini}@berkeley.edu

Introduction

Precursor

Recent work in translation [1, 2] has focused on creating large-scale multilingual transformer models which are capable of achieving state-of- the-art performance.

Key Problem

Machine translation has become very good for high resource languages in recent years but not so much for low-resource languages. Newer multilingual models are moved away from english centric modeling [1] which brought gains of > 10 BLEU

Motivation

Multilingual model [1] (M2M-100) is trained on translation task only, as opposed to T5 [3] which can be utilized for many different tasks.

The Question

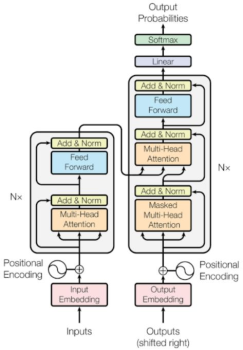
We explored the possibility of expanding the translation capability of T5 to new low-resource languages by utilizing the trained embeddings and tokenizer from multilingual model (M2M-100) as the form of transfer learning.

Models

We decided to asses the transfer learning potential between **M2M-100** and **T5** models.

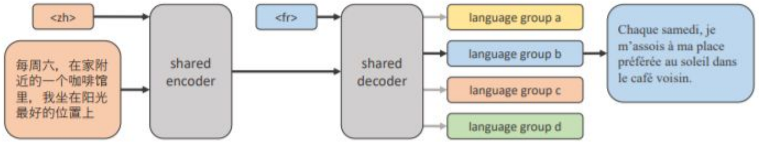
T5

T5 is an encoder decoder type transformer model which treats every text processing problem as a “text-to-text” problem allowing it to directly apply the same model, objective, training procedure and decoding process to every covered task.



M2M-100

M2M-100 is a multilingual machine translation model trained on large scale many-to-many datasets of 100 languages. Its architecture is very close to T5 and comprises of an encoder-decoder architecture but it is more versatile for translation tasks



每周六，在家附近的一个咖啡馆里，我坐在阳光最好的位置上

Chaque samedi, je m’assois à ma place préférée au soleil dans le café voisin.

Berkeley
UNIVERSITY OF CALIFORNIA

Core Hypotheses

Hypothesis 1: Embeddings as the means of transfer learning

Embeddings in the transformer models are treated as parameters and are optimized using back-propagation. This means that usage of these embeddings multilingual models can / should enable transfer learning.

Hypothesis 2: Pre-trained T5 should generalize for low-resource language translation

Usage of pre-trained T5 weights of the encoder and decoder should enhance transfer learning for an entirely new task (e.g. Translating English to Estonian)

Experiment Design & Data

Key Experiments

1. **Random Embeddings and Random Weights** (**RE-RW**): We randomly initialize the T5 weights and embedding layer and train the architecture on the 3 language pairs in all directions. This was our **baseline**.
2. **Random Embeddings and T5 Weights** (**RE-TW**): We randomly initialize the embedding layer but maintain the T5 weights and train the T5 architecture on the 3 language pairs in all directions.
3. **M2M-100 Embeddings and Random Weights** (**ME-RW**): We randomly initialize the T5 weights but maintain the M2M-100 embedding layer and train the T5 architecture on the 3 language pairs in all directions.
4. **M2M-100 Embeddings and T5 Weights** (**ME-TW**): We maintain the T5 weights and the M2M-100 embedding layer and train the T5 architecture on the 3 language pairs in all directions.

Training Data Set

Our data came from Flores-101 [4] dataset launched by FB for WMT-21 for large-scale MMT task.

We selected **English** (en), **Estonian** (et) and **Macedonian** (mk) as our target languages. We observe that the size of the training dataset varies significantly across language pairs. Macedonian (mk) has significantly less training data common with English & Estonian making it a good reference candidate for checking model performance for low resource languages.

Source / Target *	EN	ET	MK
EN	0	35.71	2.72
ET	35.71	0	3.07
MK	2.72	3.07	0

* Number of training sentences in Millions

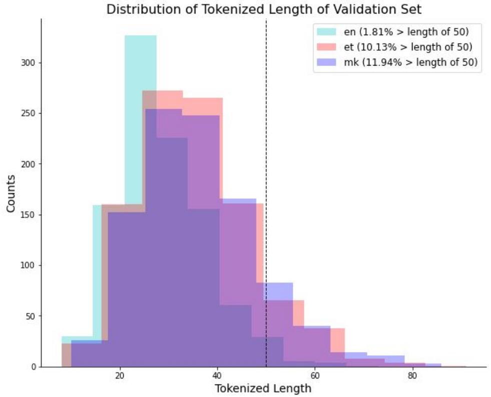
Data Processing and Clean-up

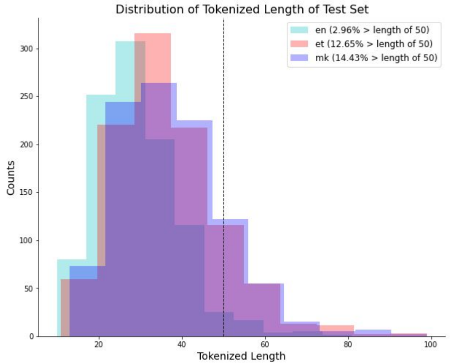
- We created 3 **training** set with 30K, 60K and 120K sentences for our experiments, each with equal number of sentence per direction

- The **validation** dataset consisted of 5,982 sentence pairs and **test** dataset had 6,072 sentence pairs.

- We employed a simple filtering criteria based on the individual lengths of the source and target sentences and the relative difference between source and target sentence sizes to filter out degenerate cases from training set.

- In order to speedup learning, we decided to set the max training length capped at 50 tokens. Looking at the distribution of tokenized length of sentences in validation and test sets, about 97% of English, 88% of the Estonian and 85% of the Macedonian sentences were unaffected by this choice, thus making it an efficient choice.





Berkeley
UNIVERSITY OF CALIFORNIA

Model Training - Hyperparameters

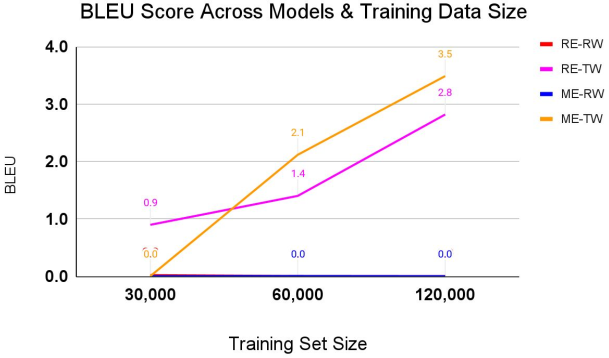
Hyperparameter	Value	Notes
Learning Rate (LR)	0.003	Constant across all model trainings
Optimizer	Adam	Constant across all model trainings
# Epochs	30	Training till convergence with early stop
GPU	Nvidia Tesla V100	
Max token length	50	

Results

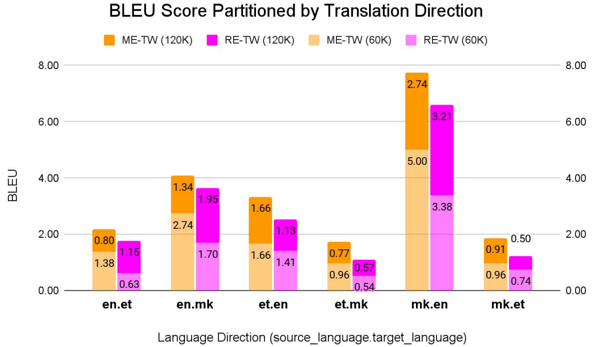
Compared to the baseline (RE-RW) model, we have evidence to suggest that maintaining T5 encoder-decoder weights results in superior performance even at low training data sizes. Augmenting the training with pre-learnt multilingual embeddings improves the performance further but this improvement is only seen with larger training set sizes.

Final BLEU score was highest for the ME-TW model with 120K sentence training size (3.5). The RE-TW model shows larger gain in BLEU score averaged across all directions (108%) compared with ME-TW model (73%) when dataset size doubles.

For low resource language pairs (et & mk), ME-TW performance is significantly better (53% avg.) than the RE-TW model for both 60K and 120K training data sizes

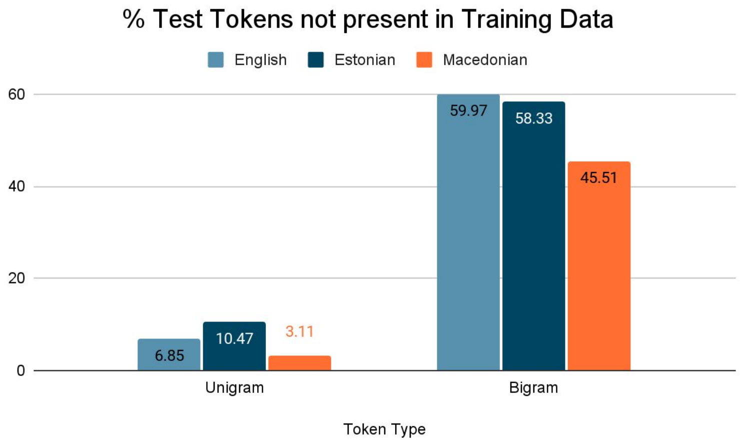


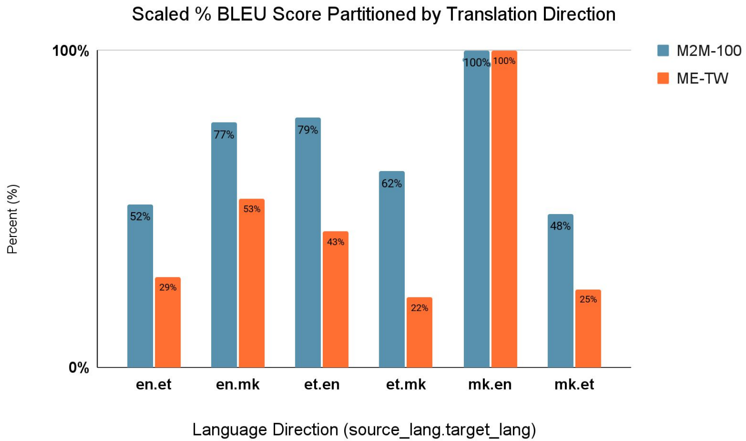
BLEU Score Across Models & Training Data Size



BLEU Score Partitioned by Translation Direction

Exploration of differential Performance





We observe a differential performance comparing mk→en translation with other directions. The two contributing factors we analysed are following:
1. A larger % of EN and ET tokens (uni and bi-grams) from test are missing in training compared to MK
2. M2M-100 model shows the similar skew in performance which indicates that our skew might be inherited.

Error Analysis - I

Source (mk): Постојат многу нешта кои треба да ги земете во предвид пред и кога патувате некаде.
Google translate ref: There are many things to consider before and when you travel somewhere.
Target (en): There are many things you have to take into consideration before and when you travel somewhere.

ME-TW translation:

- **10K training (BLEU 38.7)**: There are many things that you **need** to take into consideration before you **to me**.
- **20K training (BLEU 64.5)**: There are many things you **need** to take into consideration before and when you **do**.

Observations:

1. Google translate reference is quite close to the target sentence, however, google translate version is more succinct.
2. Both 10K and 20K models are confusing "have to" with "need to". From the perspective of translation quality, this doesn't seem like an error.
3. Model translations are first person whereas the reference translation is in third person. This is a unique artifact of the training set.
4. Both models found it hard to translate the end of sentence (i.e. "travel somewhere"). The models are not able to capture the travel context which is presented at the very end of the source sentence ("патувате некаде").

Error Analysis - II

Source (en): *He produced over 1,000 stamps for Sweden and 28 other countries.*
Google translate ref: Тој произведе над 1000 марки за Шведска и 28 други земји.
Target (mk): *Изработил преку 1000 поштенски марки за Шведска и 28 други земји.*

ME-TW translation:

- **10K training (BLEU 34.5)**: **Тој** произлезе од 1000 возач за Sweden и 28 други земји.
  - **Google translate ref**: It came from 1000 drivers for Sweden and 28 other countries.
- **20K training (BLEU 53.3)**: **Тој** произведу -ва околу 1000 бодови за Шведска и 28 други земји.
  - **Google translate ref**: It produces about 1000 points for Sweden and 28 other countries.

Observations:

1. The target macedonian sentence seems to be missing the reference for subject pronoun “He”. Interestingly, both models are able to pick this up and correct for it (“Тој”) [Appendix].
2. Both models have problems translating the word “stamp” in the source sentence. It is a possible artifact of limited exposure to the word due small training size. The 20K sentence model was able to pick-up macedonian translation of the word “Sweden”.
3. Finally, it is interesting to note that the 10K model was able to identify the correct tense of the verb (“came”) however the verb itself was incorrect. In the case of the 20K model, the verb was correct, however the tense used was wrong. It seems that model training was happening in the right direction and more training time and data might have helped the model get the correct translation.

Limitations

Max token cap

Looking at the sentence size in our validation and test set, we decided to cap the max_token_size to 50. This limits the generalizability of our exploration for longer sentences.

Training size

Our experiments included relatively small amount of paired data with a total of 120K sentence pairs covering 6 directions. Our results don't showcase the steady state performance that can be achieved by the respective models with more training time and ample data

Model explored

We explored the T5 model as the only target architecture for understanding the efficiency of transfer learning across different model types. It is possible that our results are dependent on the choice of base model architecture and hence less generalizable.

Task explored

Finally, all of our experiments were focused on the single downstream task of translation. More work needs to be done in order to ensure that our insights are generalizable across different downstream tasks such as Summarization, Classification, and Question and Answering.

Conclusion & Next Steps

We demonstrated that pre-trained encoder-decoder weights from T5 are able to generalize over untrained low-resource languages and to improve performance compared to random initialization of weights even with limited training data. Augmenting the training with pre-learnt multilingual embeddings from M2M-100 resulted in an additional boost in performance at larger training set sizes.

Multilingual embeddings significantly improved performance (+53% avg. BLEU score) on the low resource language pair (et- mk). With as little as 20K sentence pairs, the model is able to extract nuances of the languages.

We plan to test the generalizability of our hypothesis using **MBart** and **MBart-50** in future work.

References

[1] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli and Armand Joulin. Beyond English centric Multilingual Machine translation. arXiv pre- print arXiv:2010.11125, 2020.

[2] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant and Colin Raffel. mT5: A Massively Multilingual Pre- trained Text-to-text Transformer. arXiv preprint arXiv:2010.11934, 2021.

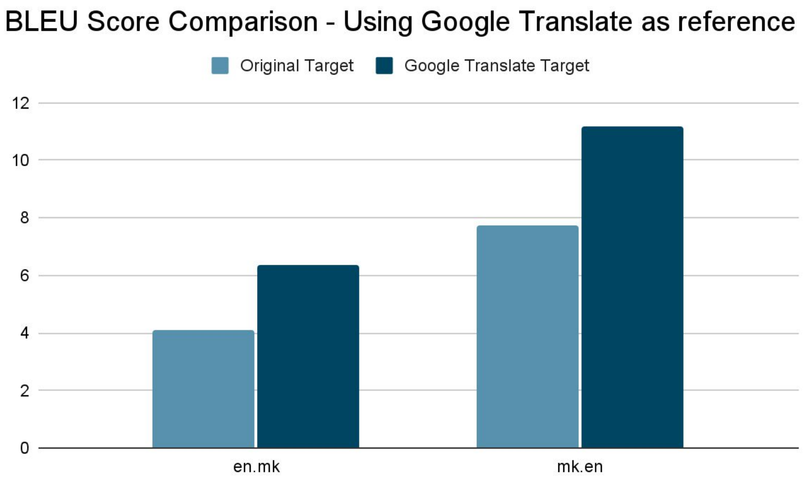
[3] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67, 2020.

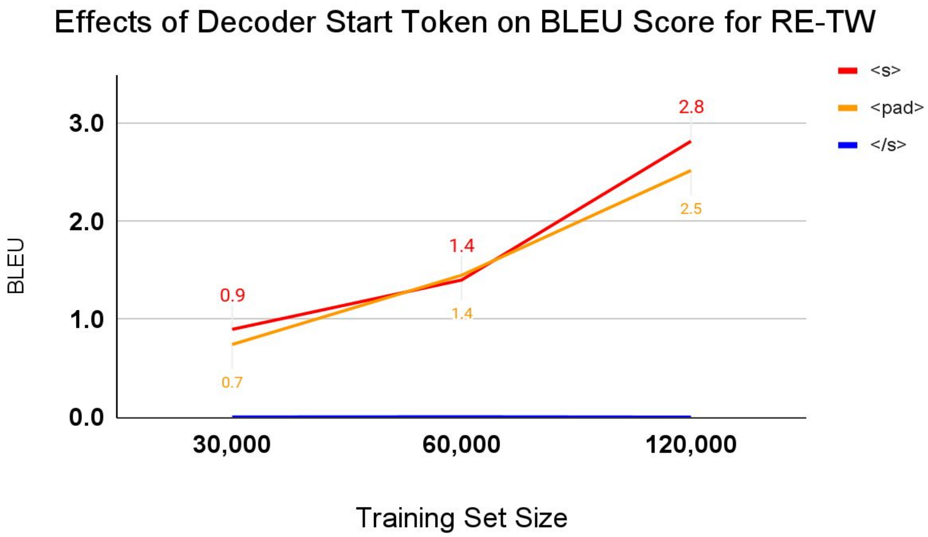
[4] EMNLP 2021: Sixth Conference on Machine Translation (WMT21)

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30, pages 5998–6008, 2017.

Appendix





Berkeley
UNIVERSITY OF CALIFORNIA