Investigating Improvement to English-Tigrinya Translation via Transfer Learning Over Varying Languages

Stanford CS224N Custom Project

Abel Dagne
Department of Computer Science
Stanford University
abeldag@stanford.edu

Sheden Andemicael
Department of Computer Science
Stanford University
sheden@stanford.edu

Abstract

In this project, we focus on improving English-Tigrinya translation through transfer learning from models pre-trained on English-Amharic, English-Arabic, English-Russian, and English-Spanish language pairs. We utilize the NLLB (No Language Left Behind) English-Tigrinya parallel corpus, which contains high-quality bitexts mined using advanced techniques such as the stopes mining library and LASER3 encoders. Our approach involves training a baseline English-Tigrinya transformer model and then applying transfer learning using weights from the aforementioned pre-trained models. We evaluate the performance of our models using both BLEU and chrF++ scores. This study aims to identify the most effective pre-trained model for transfer learning to maximize the performance of the English-Tigrinya translation model. We hypothesise that English-Amharic transfer learning will yield the greatest enhancement of English-Tigrinya translation quality because Amharic and Tigrinya are both Ge'ez-derived languages, sharing significant linguistic and structural similarities.

For this project, we have no external collaborators, no external mentor, and we are not sharing the project. Our TA mentor is Moussa Doumbomya. The breakdown of this assignemnt was 50/50 where we each researched to narrow down our problem and sourced and trained models. Abel did work to gather, clean, and preprocess the datasets while Sheden did work to write the evaluation and training code.

1 Introduction

Along with the broader field of AI, machine translation has made significant strides in recent years. However, challenges persist when it comes to low-resource languages such as Tigrinya. Tigrinya, a Semitic language spoken by millions in Eritrea and Ethiopia, lacks extensive parallel corpora necessary for training robust MT models. This limitation makes it difficult to achieve high-quality translations using standard neural machine translation techniques.

Transfer learning offers an interesting and promising solution to this problem. By leveraging knowledge from models pre-trained on other languages, perhaps those with higher resources, we can improve translation performance for low-resource languages. This approach not only enhances the quality of translations but also reduces the data and computational requirements for training effective NMT models.

In this project, we aim to improve English-Tigrinya translation through transfer learning from models pre-trained on various language pairs. Specifically, we investigate the impact of transfer learning from English-Amharic, English-Arabic, English-Russian, and English-Spanish models. Our hypothesis is that transfer learning from English-Amharic will yield the greatest improvement due to the

linguistic similarities between Amharic and Tigrinya, both of which are derived from the ancient Ge'ez language.

We utilize the NLLB English-Tigrinya parallel corpus, which contains high-quality bitexts mined using advanced techniques such as the stopes mining library and LASER3 encoders. Our methodology involves training a baseline English-Tigrinya transformer model, applying transfer learning from the selected pre-trained models, and evaluating the performance of these models using BLEU and chrF++ scores.

This study aims to identify the most effective pre-trained model for transfer learning to maximize the performance of the English-Tigrinya translation model. By exploring various language pairs and their impact on transfer learning, we hope to provide valuable insights into improving MT for low-resource languages and contribute to the broader goal of making high-quality translation accessible for all languages.

2 Related Work

Transfer learning has been a great technique in advancing machine translation for low-resource languages. Several studies have researched leveraging high-resource language pairs to improve the translation quality of low-resource languages.

One of the early works by Zoph et al. (2016) introduced a method where a parent model trained on a high-resource language pair can be fine-tuned on a low-resource language pair. This approach showed significant improvements in BLEU scores for several low-resource languages, emphasizing the effectiveness of transfer learning in neural machine translation Zoph et al. (2016).

Nguyen and Chiang (2017) extended this work by exploring transfer learning across low-resource, related languages. Their method focused on enhancing vocabulary overlap using Byte Pair Encoding, which resulted in improvements in translation quality. This study highlighted that even related low-resource languages could benefit significantly from transfer learning, with improvements of up to 4.3 BLEU points when using a stronger BPE baseline Nguyen and Chiang (2017).

Aji et al. (2020) investigated what transfer learning specifically transfers in the context of NMT. Their studies revealed that word embeddings play a crucial role, especially when they are properly aligned. They found that while transfer learning can be effective without embeddings, the results are suboptimal. Interestingly, even using randomly generated sequences as parent languages yielded noticeable gains, though smaller compared to real languages. This study underscored the importance of embeddings and suggested that transfer learning can eliminate the need for a warm-up phase when training transformer models in high-resource language pairs Aji et al. (2020).

Further experiments by researchers demonstrated that the choice of the parent language significantly impacts the performance of the transfer learning model. For instance, using French as a parent language for low-resource languages like Spanish showed better BLEU scores compared to using German, likely due to the linguistic similarities between French and Spanish. This indicates that selecting a linguistically similar parent language can optimize the benefits of transfer learning Aji et al. (2020).

We found studies that have applied these techniques to various language pairs, including Hausa, Turkish, Uzbek, and Urdu, showing consistent improvements in BLEU scores when using transfer learning. The improvements varied across languages, but the overall trend supported the robustness of transfer learning in enhancing translation quality for low-resource languages Zoph et al. (2016); Nguyen and Chiang (2017); Aji et al. (2020).

3 Approach

1. **Baseline Model:**
    - Train a baseline English-Tigrinya transformer model. Using the M2M100ForConditionalGeneration architecture from Hugging Face, we developed our baseline NMT model using the hyperparameters as outlined in Adhanom (2021).
    - Record its performance using the BLEU score and chrF++ score.

2. **Select Pre-trained Models:**
    - Identify strong transformer models that are pre-trained on English-Amharic, English-Arabic, English-Russian, and English-Spanish language pairs.
3. **Apply Transfer Learning:**
    - Initialize the English-Tigrinya model with pre-trained weights from selected models.
    - Fine-tune each model on the English-Tigrinya dataset using the same hyperparameters for consistency.
4. **Evaluate Performance:**
    - Compare the BLEU and chrF++ scores of the fine-tuned models against the baseline.
    - Analyze the improvements and differences in performance.

4 Chosen Models for Transfer Learning

In our approach to enhance English-Tigrinya translation via transfer learning, we have selected four pre-trained transformer models. These models were chosen based on their linguistic relevance and the availability of high-quality pre-trained models. Below, we detail the rationale and information about each selected model:

1. **English-Amharic Model:** Pre-trained transformer model
   Link: `https://huggingface.co/Atnafu/English-Amharic-MT`
   Rationale: Amharic and Tigrinya both belong to the Semitic language family and share significant linguistic and structural similarities. The shared heritage from Ge'ez means that syntactic, morphological, and lexical features may be closely related, making it a strong candidate for transfer learning to Tigrinya.

2. **English-Arabic Model:** Pre-trained transformer model trained on data from Tiedemann and Thottingal (2020)
   Link: `https://huggingface.co/khalidalt/m2m100_418M-finetuned-en-to-ar`
   Rationale: Arabic is another Semitic language and, while not as closely related to Tigrinya as Amharic, it still shares some structural and grammatical features. Additionally, Arabic has a vast amount of resources and pre-trained models available, making it a suitable candidate for transfer learning.

3. **English-Spanish Model:** Pre-trained transformer model trained on data from Tiedemann and Thottingal (2020)
   Link: `https://huggingface.co/cartesinus/iva_mt_wslot-m2m100_418M-en-es`
   Rationale: Spanish, although not related to Tigrinya, is a major language with a vast amount of high-quality pre-trained models available. Using a well-resourced language like Spanish can help understand the impact of transfer learning from a high-resource language, providing valuable insights into the transfer learning process.

4. **English-Russian Model:** Pre-trained transformer model trained on data from Tiedemann and Thottingal (2020)
   Link: `https://huggingface.co/kazandaev/m2m100_418M-finetuned-en-ru`
   Rationale: Russian is not linguistically related to Tigrinya, but it is included to test the impact of transfer learning from a linguistically distant language. Russian has extensive resources and high-quality pre-trained models, which can be used to test the robustness of transfer learning across unrelated language pairs.

By using these diverse models, we aim to gain a comprehensive understanding of how linguistic similarity and resource availability impact the effectiveness of transfer learning for improving English-Tigrinya translation.

5 Experiments

5.1 Data

The dataset we are using for this project is the NLLB English-Tigrinya parallel corpus. This dataset is part of the NLLB (No Language Left Behind) project by Meta AI, which aims to create high-quality parallel corpora for various language pairs. This dataset was created based on metadata for mined bitext released by Meta AI. It contains bitext for 148 English-centric and 1465 non-English-centric language pairs using the stopes mining library and the LASER3 encoders.Fan et al. (2021) Schwenk et al. (2021). The task associated with this dataset is to train our transformer model for Neural Machine Translation from English to Tigrinya.

The FLORES+ evaluation set is used to assess the performance of neural machine translation models, particularly for low-resource language pairs such as English-Tigrinya. Developed as part of the NLLB initiative, FLORES+ provides a standardized benchmark for evaluating translation quality across numerous languages AI (2022). The dataset consists of manually translated sentences, ensuring high-quality, human-generated references for model evaluation. By leveraging the FLORES+ evaluation set, we can rigorously test the accuracy and fluency of our trained transformer model. This evaluation set allows us to perform a comprehensive analysis of the translation performance, facilitating a deeper understanding of the strengths and weaknesses of our model in translating between English and Tigrinya.

Preprocessing Steps:

1. **Text Normalization:** Normalize the text by converting all characters to lowercase and removing any extraneous spaces or punctuation that does not contribute to the meaning.
2. **Tokenization:** Tokenize the text into subwords using the SentencePiece library, which helps in handling rare words and improving translation quality.
3. **Vocabulary Generation:** Generate a shared vocabulary for both the source and target languages, ensuring that the model can handle words and phrases across both languages effectively.
4. **Dataset Splitting:** Split the dataset into training and validation sets. The splits are performed in such a way to maintain the distribution of sentences across these sets.
5. **Binning and Bucketing:** Organize the data into buckets of similar lengths to minimize padding and improve computational efficiency during training.

5.2 Evaluation method

- **BLEU Score:** The BLEU score was used which measures the accuracy of the translated sentences by comparing them to reference translations. However, BLEU can be limited in evaluating translations of morphologically rich languages.
- **chrF++ Score:** To address the limitations of BLEU, we also use the chrF++ score. While chrF evaluates translations based on character n-grams, chrF++ extends this by incorporating word n-grams, making it more suitable for morphologically rich languages like Tigrinya. This provides a more comprehensive assessment of translation quality by capturing both character-level and word-level information.

5.3 Experimental details

- **Model configurations:** All pre-trained models selected for our experiment use the M2M100ForConditionalGeneration transformer architecture. These models has 16 attention heads. Input embeddings have a dimensionality of 1024 and the hidden layers have a dimensionality of 4096.
- **Hyperparameters:** We use a learning rate of $1 \cdot 10^{-4}$ with a rate decay of $1 \cdot 10^{-5}$. Additionally, we use a per-device batch size of 16 and train the models for 3 epochs.
- **Training time:** Each model took around 7 hours to fine-tune on 250,000 sentence pairs.

5.4 Results

We present the quantitative results of our experiments in the table below. The BLEU and chrF++ scores for the baseline and transfer learning models are reported:

Model	BLEU Score	chrF++ Score
Baseline Model	2.7982	18.4549
En-Am Parent Model	**3.3080**	**19.8116**
En-Ar Parent Model	3.1309	18.7770
En-Es Parent Model	3.0758	18.9326
En-Ru Parent Model	2.8320	18.7785

Table 1: Comparison of BLEU and chrF++ scores for baseline and transfer learning models on the English-Tigrinya translation pair

5.5

Our baseline scores were worse than we expected. For our baseline English-Tigrinya model, we were unable to achive the SOTA scores on the FLORES+ evaluation set. Despite our baseline scores being lower than the state-of-the-art, our experiment remains valid and provides valuable insights into the effectiveness of transfer learning for English-Tigrinya translation. The primary objective of our study was to evaluate the relative improvements achieved through transfer learning from various pre-trained models rather than to establish absolute state-of-the-art performance. Thus, while our baseline scores are lower, the relative improvements observed validate our experimental approach and contribute to a better understanding of transfer learning's efficacy in the context of low-resource language translation.

6 Analysis

The results indicate that transfer learning significantly improves the performance of the English-Tigrinya translation model when using the English-Amharic parent model, as evidenced by the substantial increase in both BLEU and chrF++ scores. This supports our hypothesis that linguistic similarities between Amharic and Tigrinya can enhance translation quality. However, the models with other parent languages (Arabic, Spanish, and Hebrew) showed minimal improvements, suggesting that the effectiveness of transfer learning is closely tied to the linguistic and structural similarities between the parent and target languages.

It is important to note that BLEU performs worse at accurately measuring the quality of translations for morphologically complex languages. chrf++ works as a better measure for quickly computing the quality of Tigrinya translations. Although a sentence could accurately convey the same meaning as the source sentence, it may still score low on BLEU. This serves as an explanation for the low BLEU scores relative to other language pairs. After randomly selecting and examining translations from each of the English-Tigrinya models, specifically, the child model fine-tuned on the English-Amharic parent model, we were able to verify the quality of these translations with a native Tigrinya speaker. Some translations are replicated in table 2 below.

7 Conclusion

In this study, we aimed to improve English-Tigrinya translation through transfer learning using models pre-trained on English-Amharic, English-Arabic, English-Russian, and English-Spanish language pairs. Our results demonstrated that transfer learning can significantly enhance translation performance, especially when leveraging a linguistically similar parent language like Amharic. The substantial increase in both BLEU and chrF++ scores for the English-Amharic parent model supports our hypothesis that linguistic similarities play a crucial role in the effectiveness of transfer learning.

While our baseline scores were lower than the state-of-the-art, the relative improvements achieved through transfer learning validate our experimental approach and provide valuable insights into the

Model	Sentences
Source	So, it is likely that the notation was added simply as a label.
Reference	ስለዚ እቲ ጽሑፍ ከም ምልክት ጥራይ ተወሰኸ ኪኸውን ይኽእል እዩ ።
Baseline Model	ስለዚ፡ እቲ ሓበረታ ከም መልእኽቲ ጥራይ እዩ ተወሲዱ ዚኸውን።
En-Am Parent Model	ስለዝኾነ ድማ እቲ መግለጺ ከም ሓደ ኣዋጅ ጥራይ ክኸውን ይኽእል እዩ።
En-Ar Parent Model	ስለዚ፡ እቲ ሓሳባት ከም ሓንቲ ጽሕፍቲ ጥራይ እዩ ተወሳኺ ኪኸውን ዚኽእል።
En-Es Parent Model	በዚ ምኽንያት እዚ ድማ እቲ ሓበረታ ከም መልእኽቲ ጥራይ እዩ ተወሲዱ ዘሎ።
En-Ru Parent Model	ስለዚ፡ እቲ ሓበረታ ከም ምልክት ጥራይ እዩ ተወሲዱ።

Table 2: English to Tigrinya translations by each of the four models on the same randomly selected sentence from the FLORES+ evaluation set

potential of transfer learning for low-resource languages like Tigrinya. The minimal improvements observed with other parent languages highlight the importance of linguistic and structural similarities in optimizing transfer learning benefits.

Future Work. For future work, several avenues can be explored to further enhance the performance of English-Tigrinya translation:

- **Full NLLB Dataset:** Training on the full NLLB dataset could provide more comprehensive data coverage, potentially leading to better translation quality.
- **Additional Semitic Language Pairs:** Exploring transfer learning with other Semitic languages such as Hebrew could offer further insights into the impact of linguistic similarities on translation performance.
- **Warm Transfer Learning:** Implementing warm transfer learning approaches, where the model is first pre-trained on a large, related dataset before fine-tuning on the target language pair, could enhance the model's ability to generalize and improve translation accuracy. This method has shown promising results in various language pairs, as demonstrated by Thompson et al. (2019).

These directions could help in achieving state-of-the-art performance and contribute to the development of more robust and accurate translation systems for low-resource languages like Tigrinya.

8 Ethics Statement

Developing machine translation systems for languages like Tigrinya, while promoting linguistic inclusivity, poses several ethical challenges and potential societal risks. Firstly, there is the risk of cultural misrepresentation or bias in translation outputs. Machine translation systems may inadvertently perpetuate stereotypes or misinterpret cultural nuances, especially for low-resourced languages where training data may not be sufficiently diverse or balanced. This could lead to misunderstandings that affect cross-cultural interactions or perpetuate cultural biases.

Furthermore, transfer learning, while beneficial for improving translation quality, may inherit biases from the pre-trained models on which it is based. If these parent models contain biased data, the biases could be transferred and potentially amplified in the target low-resource language model. This

could exacerbate the issue of cultural misrepresentation and perpetuate systemic biases present in more widely spoken languages.

In the case of English-Tigrinya translation specifically, there is concern about the machine translation system inheriting biases from the Amharic model. Currently, in Ethiopia and Eritrea, tensions are high between ethnic groups, and there is significant division between Amharic and Tigrinya speakers. Imposing Amharic biases on Tigrinya text could contribute to escalating these tensions.

Mitigation Strategies. To combat cultural misrepresentation and bias, it is crucial to involve native speakers in the dataset curation and model evaluation process. This involvement ensures that the training data is diverse and culturally representative and that the translation outputs are reviewed for cultural accuracy and sensitivity. For instance, native speakers can identify Amharic biases present in translated Tigrinya text and address the transferred biases. Additionally, employing techniques like debiasing in the model training process can further reduce the risk of cultural bias. Regular audits of the models for bias and continuous updates with new, diverse data can help mitigate these issues.

It is also important to critically evaluate the parent models used for transfer learning for biases before using them for training low-resource language models. Techniques such as adversarial training and bias detection can be employed to identify and mitigate biases in these models. Additionally, using a diverse set of parent languages can help balance the transfer learning process and reduce the risk of amplifying specific biases.

These measures would help build a translation system that is ethically responsible, culturally sensitive, and respectful of user privacy, contributing positively to the field of machine translation and the communities it serves.

References

Isayas Adhanom. 2021. A first look into neural machine translation for tigrinya. *ResearchGate*.

Meta AI. 2022. No language left behind: Scaling human-centered machine translation. *Meta*.

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. 2020. In neural machine translation, what does transfer learning transfer? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7701–7710, Online. Association for Computational Linguistics.

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. 2021. Beyond english-centric multilingual machine translation. *arXiv preprint arXiv:2109.01707*.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:2107.07663*.

Brian Thompson, Matt Post, and Mohit Bansal. 2019. Warm-starting neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 169–181.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 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.