Adapting High-resource NMT Models to Translate Low-resource Related Languages

Group 1: Lea, Mohit, Jardenna, Mehdi

Introduction: Situation

- SOTA approaches for NMT are **data intensive**
- **Low resource** data on **Brazilian Portuguese to English** pairs ⇒ hard to train

Introduction: Motivation

But we know that :

- European Portuguese (EP) and Brazilian Portuguese (BP) are **similar**
- EP-English data is high resource ⇒ can build good EP-English models
- Less data ⇒ can train only few parameters

Do you see the pattern?

Introduction - Our approaches

1. BP-English zero shot translation using EP-English model
2. Fine tuning the whole EP-English model on BP-English data
3. Freeze the EP-English model and add adaptation layers on top of it

How similar are EP and BP?

History:

- BP origin: Colonization of the South America
- Geographical distance ⇒ differences

Pronunciation: Mostly similar except vowels and ‘s’

Accents: BP is phonetically pleasing to the ear and has strong cadence

Formal vs Informal: EP is more formal

Grammar and Spelling:

- Some differences in spelling: “receção” (EP) vs “recepção.” (BP), “you”: “você” / “tu” (EP) vs “tu” (BP)
- BP: converting some nouns into verbs + English influence

Effects Involved

Catastrophic forgetting occurs when model is trained sequentially on multiple tasks.

Why do we care about it?

Here:

- It will occur when fine-tuned
- It should be less when adapted

Other benefits of Adaptation over Fine-tuning

1. task-specific layer-wise representation learning
2. small, scalable, sharable,
3. reproducibility
4. modularity of representations
5. non-interfering composition of information

Background - Finetuning

Popular way of applying transfer learning

1. Copy weights from pre-trained model for task A
2. Finetune by continue training the model on downstream task B
   - Original Models weights are changed
   - Take into account the characteristics of domain data of task B

Background - Catastrophic Forgetting

Catastrophic forgetting was first observed by McCloskey et al. in 1989

- Continual learning often results in erasure of previous knowledge
- plasticity - stability dilemma
    - tuning parameters to most optimum learning algorithm

      -> sensitive to distributional shift (plasticity)
    - maintaining past knowledge, to reduce forgetting (stability)

Background - Adaptation Layers

Introduced in 2019 by Houlsby et. al as adapter modules.

Alternative to Finetuning.

1. Take a pre-trained model for task A
2. Freeze pre-trained model
3. Add Adaptation Layers between the Layers of the Pretrained Model
4. Train (only the Adaptation Layers) on data for task B

Background - Adaptation Layers

1. task-specific layer-wise representation learning
2. small, scalable, sharable,
3. reproducibility
4. modularity of representations
5. non-interfering composition of information

Diving into the implementation

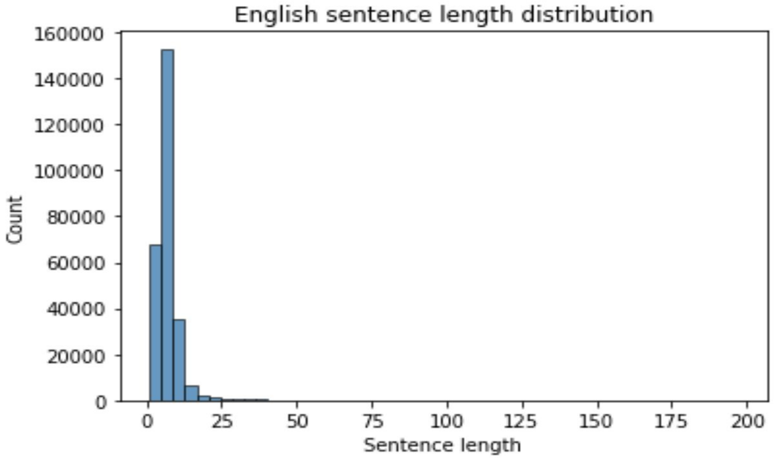
Dataset (EN - eu-PT)

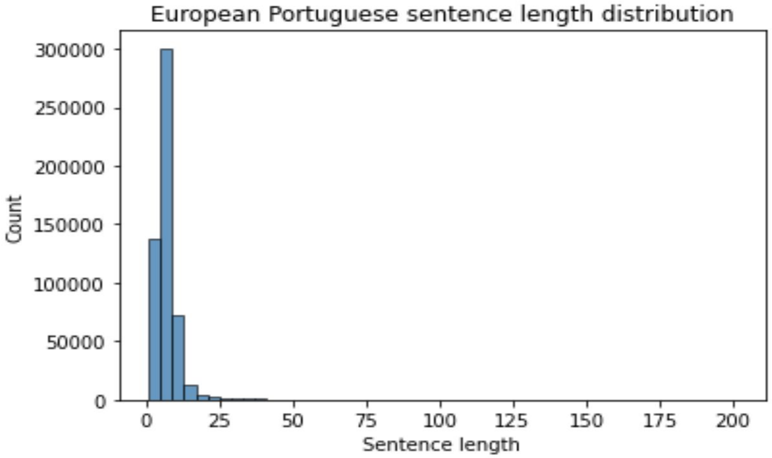
English

- vocab size: 180 K
- nr sequences: 270 K
- avr. sentence length: 6.69

European Portuguese

- vocab size: 180 K
- nr sequences 270 K
- avr. sentence length: 6.71





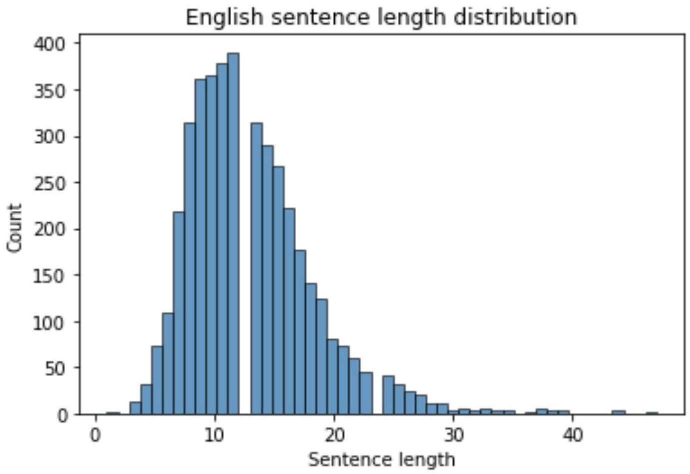
[Pirá] Dataset (EN - br-PT)

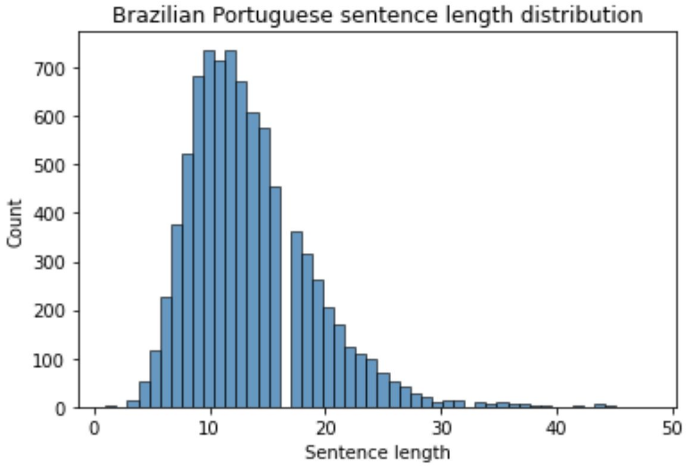
English

- vocab size: 55 K
- nr sequences: 9 K
- avr. sentence length: 13.01

Brazilian Portuguese

- vocab size: 59 K
- nr sequences 9 K
- avr. sentence length: 13.86

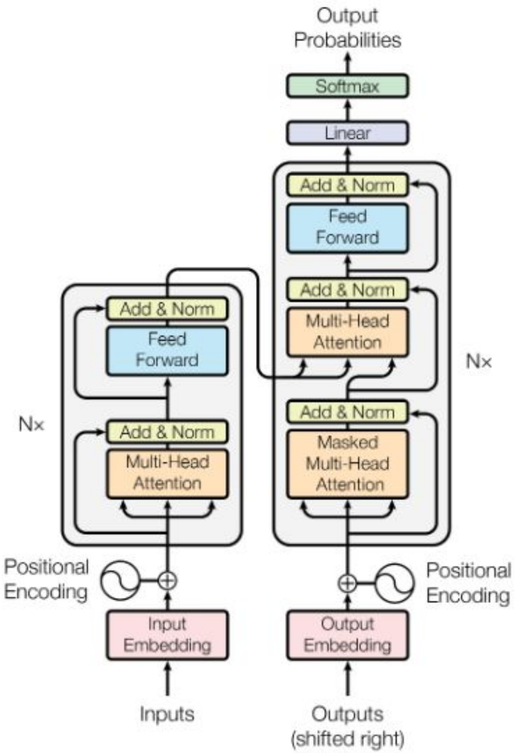




Method - The Base Model

Transformer, implemented with the same architecture as the original Transformer paper:

- Reduced Size:
    - Feedforward Layers in Encoder and Decoder: 8 -> 4
    - Key dimension size : 512 -> 128
    - Feedforward Layers: (512, 2048) -> 5(128, 512)
    - Attention heads: 8 -> 8
- Training: 20 Epochs



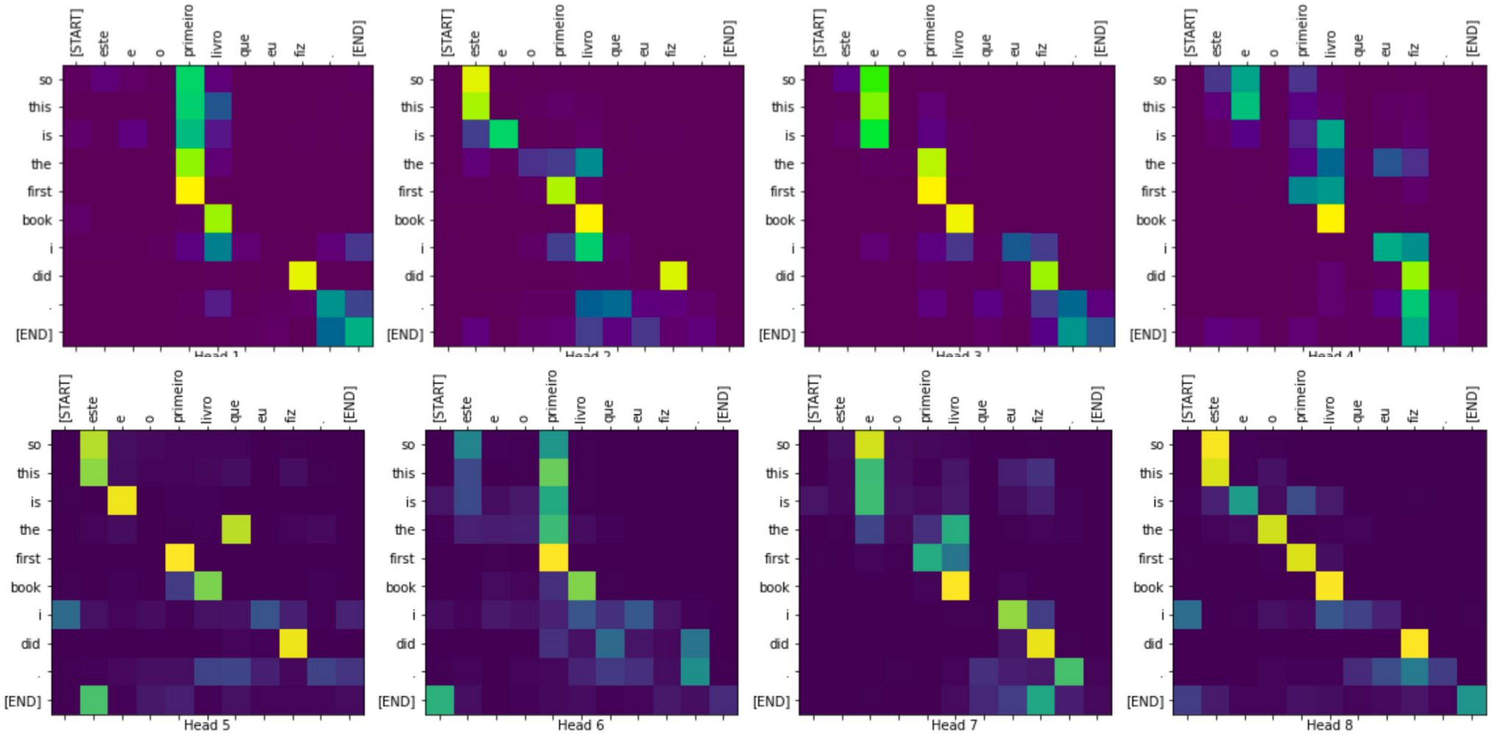
Qualitative Analysis

Input (euPT): este é um problema que temos que resolver.

Prediction: so this is a problem we have to solve.

Groundtruth: this is a problem we have to solve

Attention weights



Method - Adaptive Model

The Adaptive Model

- Base Model Transformer trained on *european Portuguese - English Dataset*
- Froze all layers
- Added 2 Dense Layers before the final Layer
- Trained the new model on *brazilian Portuguese - English Dataset*

Idea

- Forcing Model to learn european Portuguese and using this knowledge to learn brazilian Portuguese
- Even if catastrophic forgetting happens, we can recover the base model
- Enabling model to learn two close languages at the same time, without further Input (such as tokens)

Method - Finetuned Model

- Used Base Model Transformer Pretrained on *european Portuguese - English* Dataset
- Set all parameters to trainable
- Finetuned on brazilian Portuguese Dataset
    - Trained for 20 Epochs

Experiments

1. Design Transformer and train on eu-PT data

Experiments

1. Design Base Model and train on eu-PT data
    a. Evaluate Base Model on euPT
    b. Zero-Shot Performance of Base Model on brPT
        i. establishes how similar the languages are and how transferable the learned knowledge is
2. Finetune Base Model to brPT
    a. Evaluate Finetuned Model on euP
    b. Evaluate Finetuned Model on brPT
3. Add Adaptive Layers to Base Model, freeze Parameters of Base Model
    a. Evaluate Model on euPT
    b. Evaluate Model on brPT

Results

BLEU Scores	eu-PT	br-PT
Base Model	20.00	10.98
Finetuned Model	4.53	30.50
Adaptive Model	13.87	13.16

Discussion

- The Base Model

Conclusion

As expected, catastrophic forgetting reduced when model is adapted compared to fine-tuned model

Could produce better performance if:

1. different tokenizers were used
2. feed forward layers in embeddings, encoder or decoder are adapted
3. trained for more epochs (>20)

Questions?

Thank You For Your Attention!