Quebec
Artificial
Intelligence
Institute

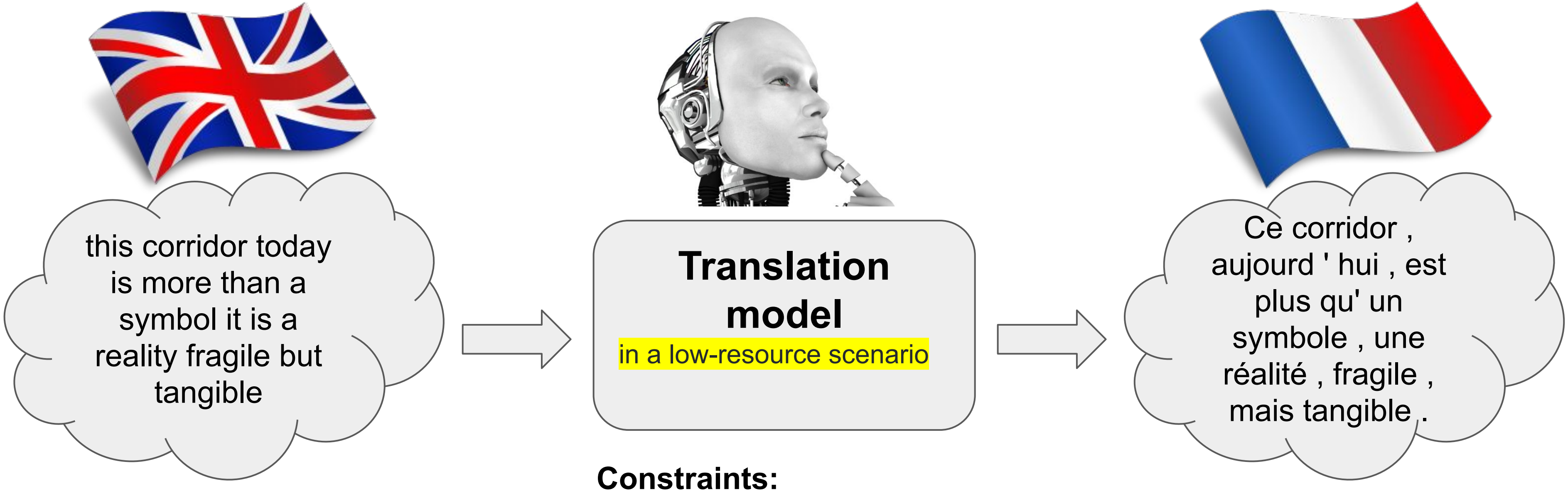
Mila

IFT6759 - Project 2

Low Resource Machine Translation

Alex Peplowski
Harmanpreet Singh
Marc-Antoine Provost
Mohammed Loukili

Project Description



this corridor today is more than a symbol it is a reality fragile but tangible

Translation model

in a low-resource scenario

Ce corridor , aujourd ' hui , est plus qu' un symbole , une réalité , fragile , mais tangible .

Constraints:

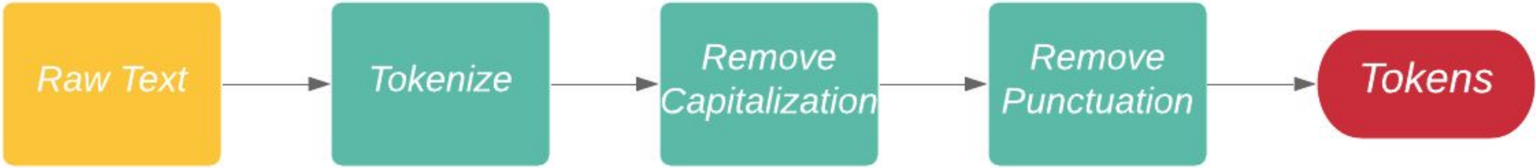
- No external data!
- No pre-trained models!
- No pre-trained embeddings!

Dataset Description

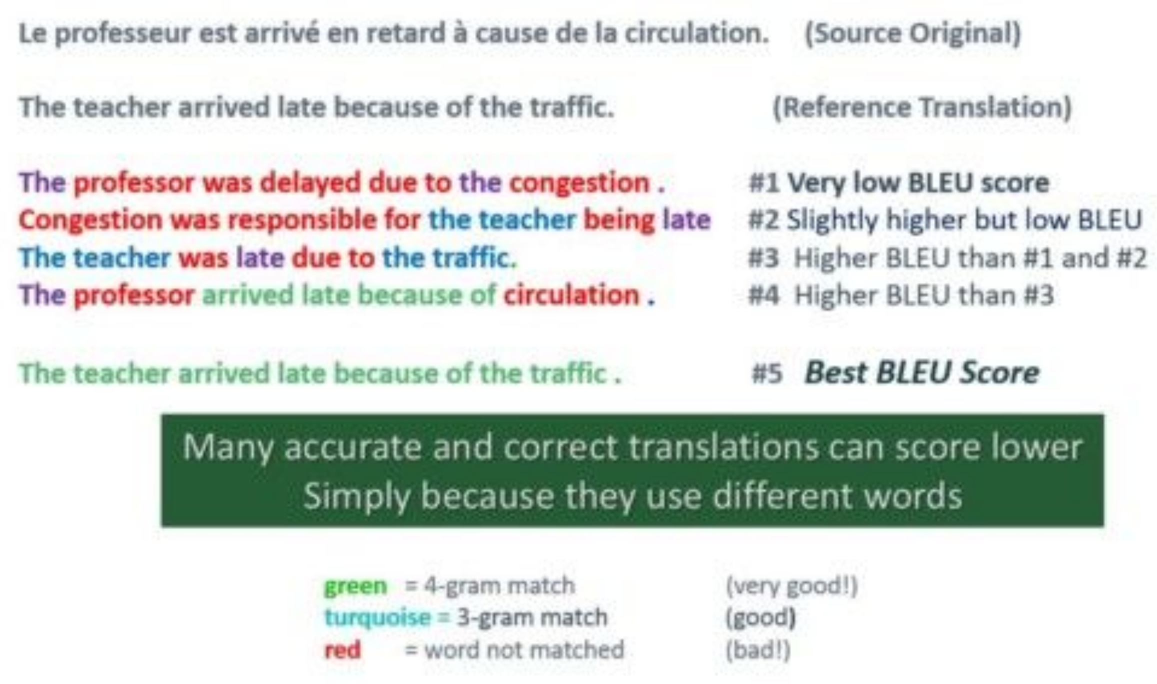
Table: Raw Data Summary

	EN		FR	
	aligned	unaligned	aligned	unaligned
Size	11k	474k	11k	474k
Tokenized?	Yes	-	Yes	-
Capitalization?	-	Yes	Yes	Yes
Punctuation?	-	Yes	Yes	Yes

Provided functions:



Evaluation Metric: BLEU Score



Le professeur est arrivé en retard à cause de la circulation. (Source Original)

The teacher arrived late because of the traffic. (Reference Translation)

The professor was delayed due to the congestion . #1 Very low BLEU score
Congestion was responsible for the teacher being late #2 Slightly higher but low BLEU
The teacher was late due to the traffic. #3 Higher BLEU than #1 and #2
The professor arrived late because of circulation . #4 Higher BLEU than #3

The teacher arrived late because of the traffic . #5 *Best BLEU Score*

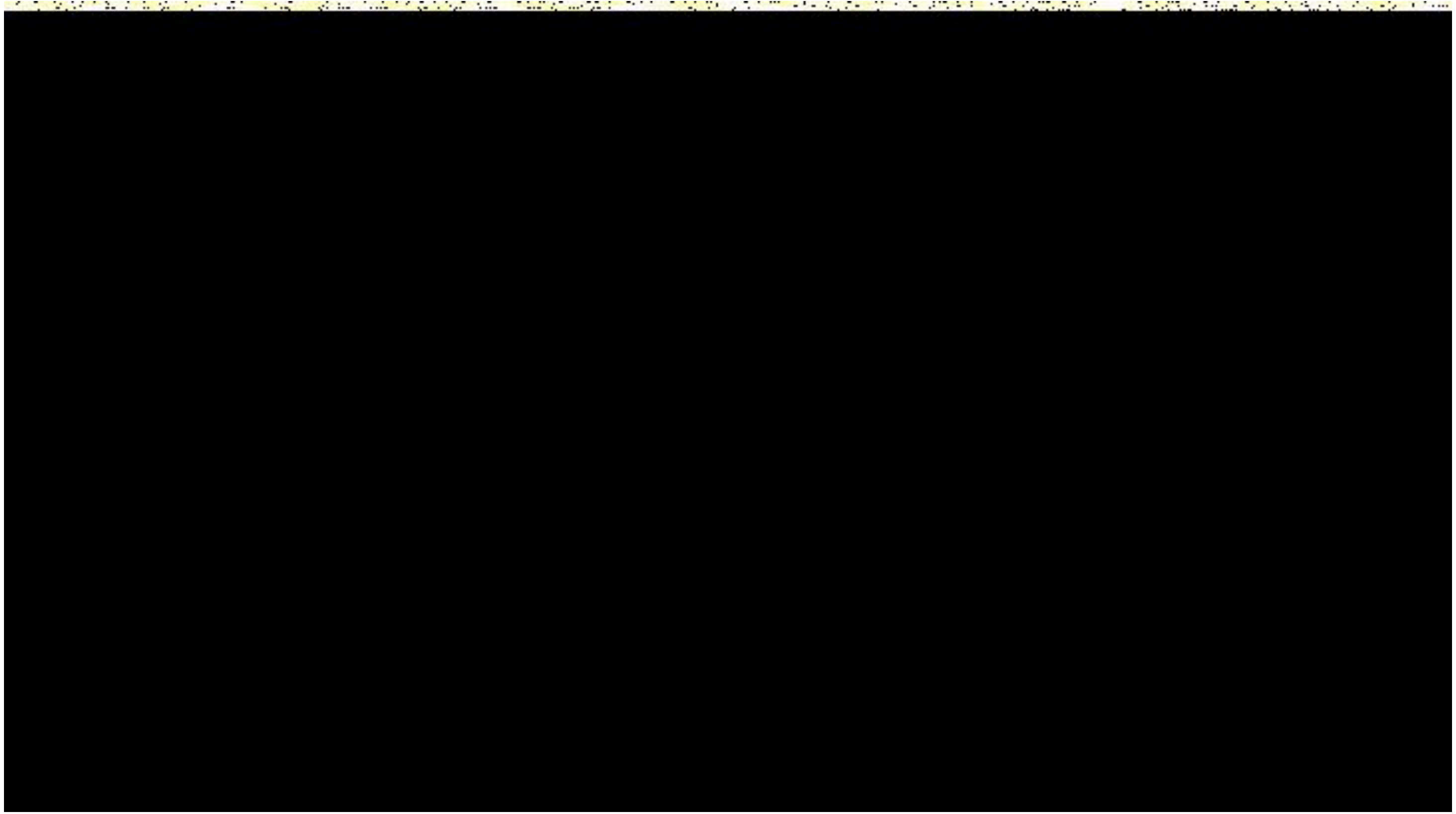
Many accurate and correct translations can score lower
Simply because they use different words

green = 4-gram match (very good!)
turquoise = 3-gram match (good)
red = word not matched (bad!)

Source: https://www.sdl.com/blog/understanding-mt-quality-bleu-scores.html

Data Ingestion/Preprocessing

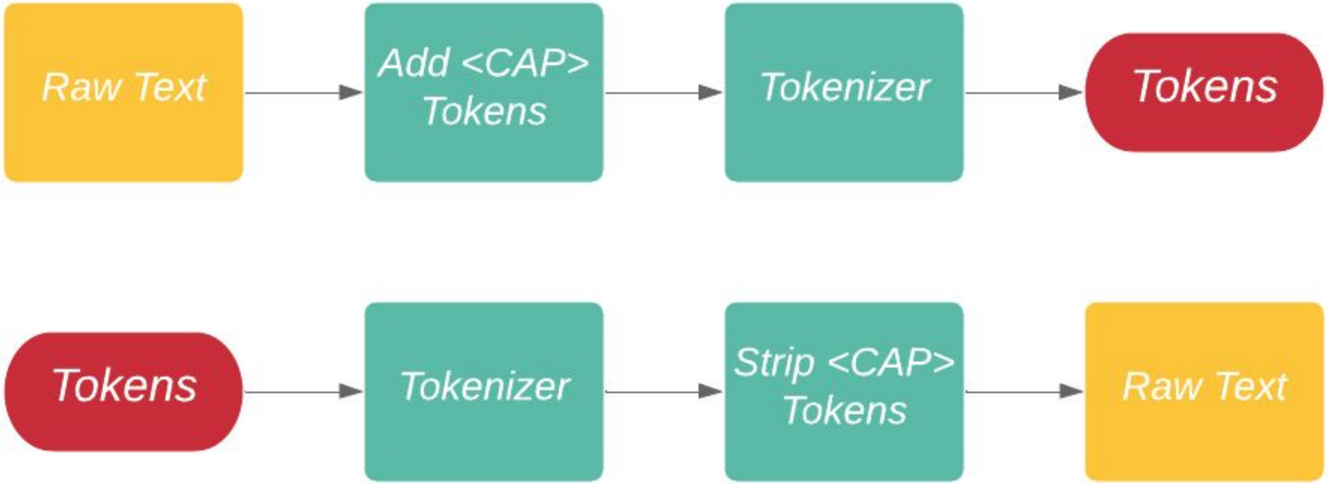
Byte Pair Encoding (BPE) Tokenizer



Source: https://towardsdatascience.com/byte-pair-encoding-the-dark-horse-of-modern-nlp-eb36c7df4f10

Capitalization

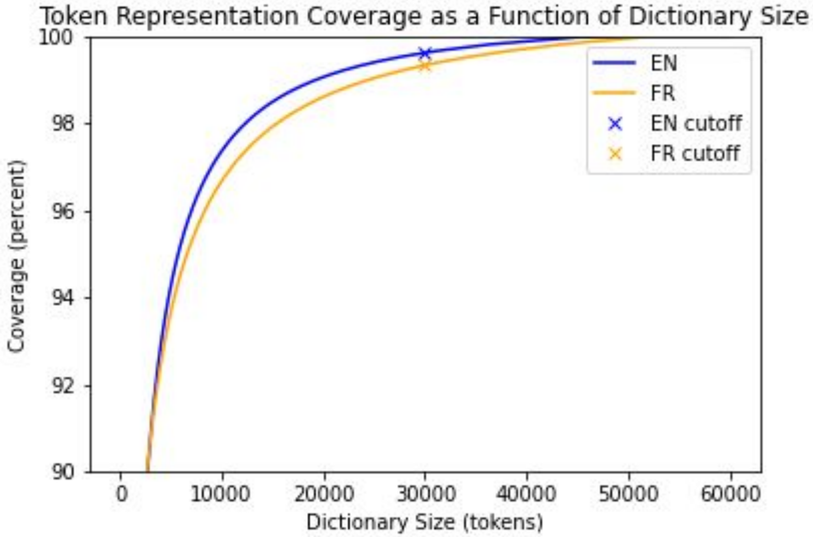
How to handle capitalization requirement?

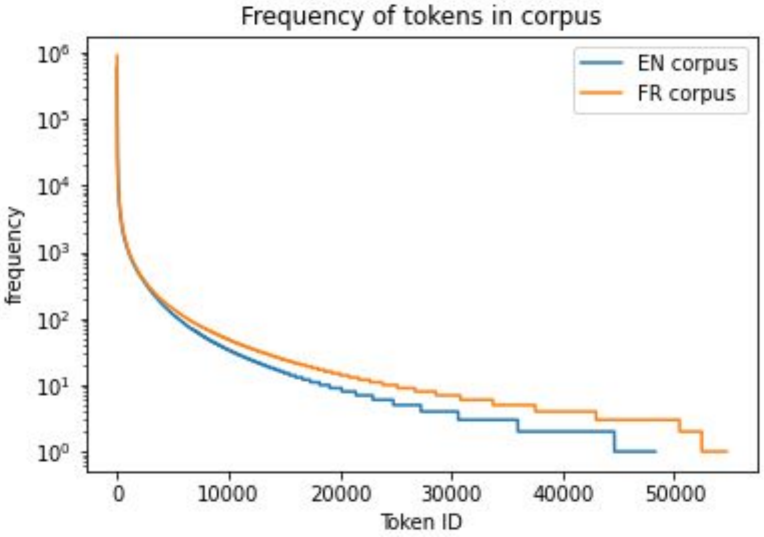


[Cat] ↔ [<CAP>, cat]

Vocabulary Size

How many unique tokens should be retained?



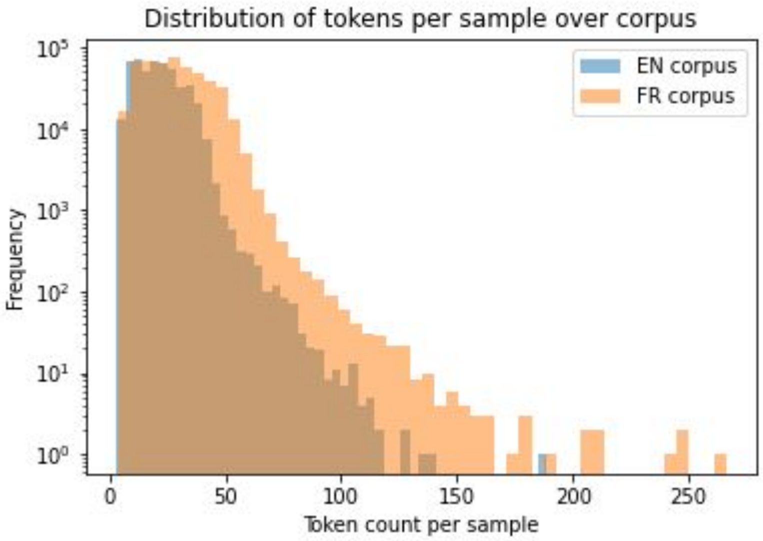


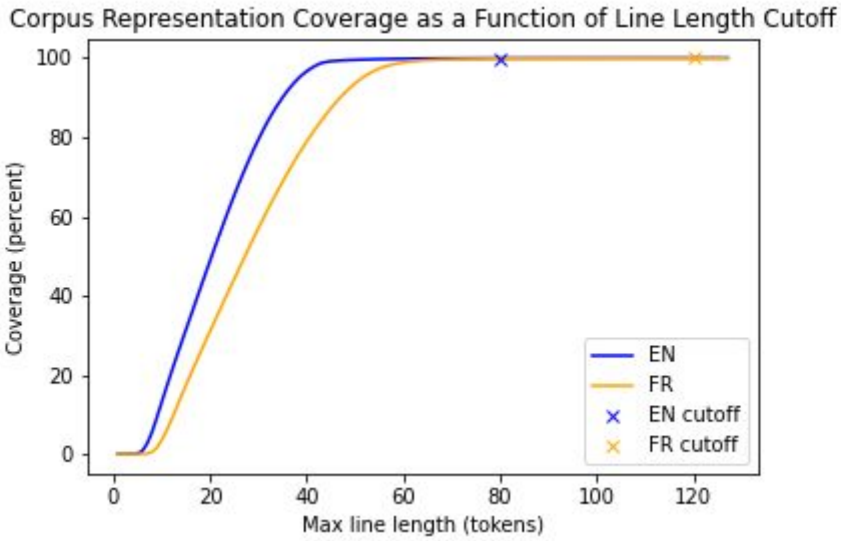
EN Coverage (size 30k): 99.6%

FR Coverage (size 30k): 99.3%

Sequence Length Upper Bound

How many tokens to keep per sentence?





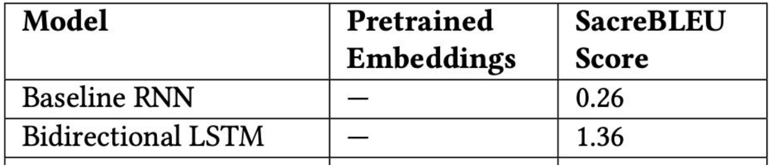
EN Coverage (length 80): 99.97%

FR Coverage (length 120): 99.98%

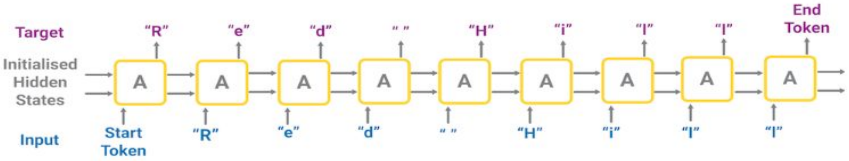
Models + Architectures

RNN and Bidirectional LSTM

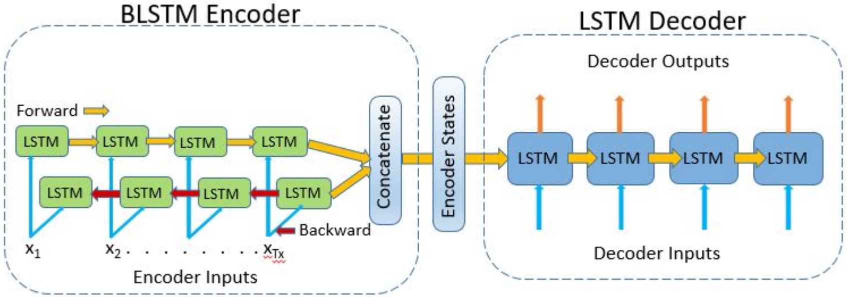
+ ELMo



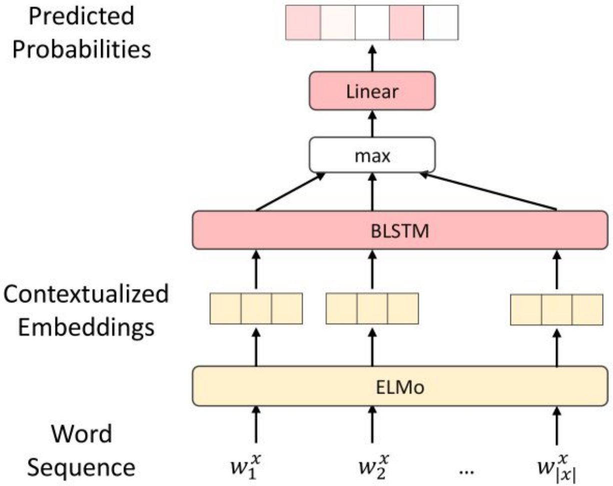
Model	Pretrained Embeddings	SacreBLEU Score
Baseline RNN	—	0.26
Bidirectional LSTM	—	1.36



RNN Encoder-Decoder



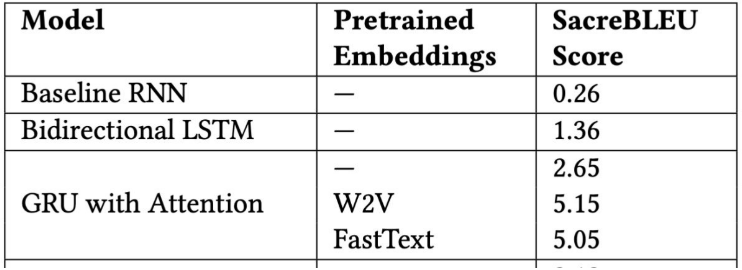
Bidirectional LSTM



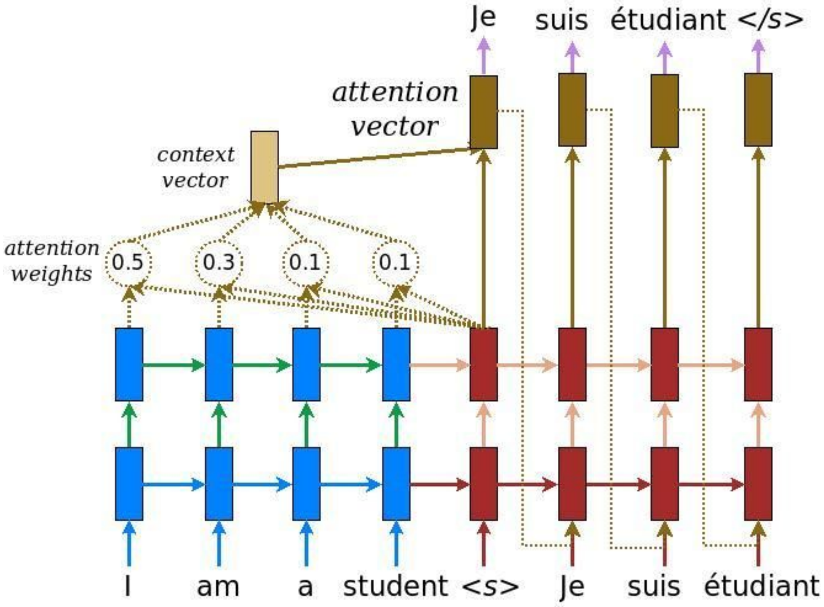
Sources from: https://towardsdatascience.com/time-series-forecasting-with-deep-stacked-unidirectional-and-bidirectional-lstms-de7c099bd918
https://www.groundai.com/project/learning-asr-robust-contextualized-embeddings-for-spoken-language-understanding/

GRU with Attention

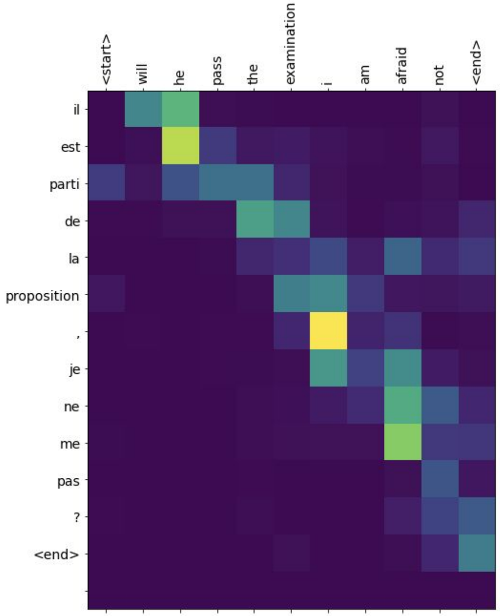
+ w2v / fasttext



Model	Pretrained Embeddings	SacreBLEU Score
Baseline RNN	—	0.26
Bidirectional LSTM	—	1.36
GRU with Attention	—	2.65
	W2V	5.15
	FastText	5.05



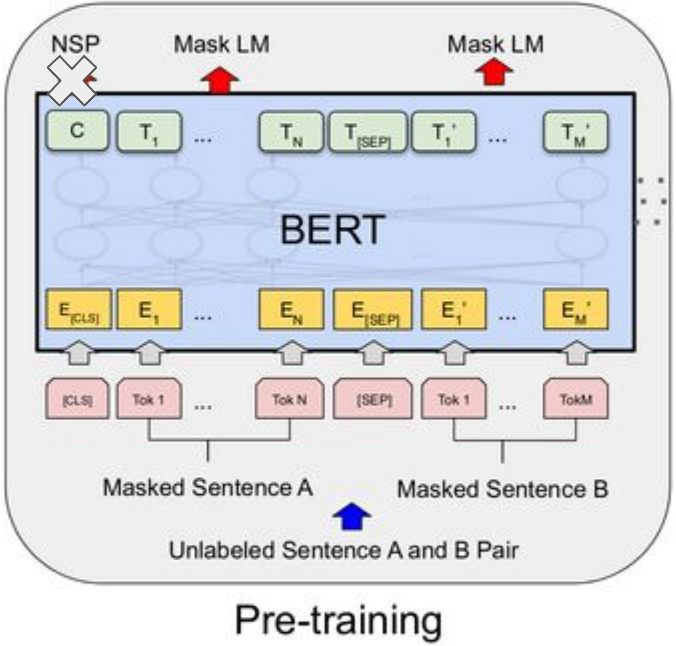
Encoder-Decoder model with attention

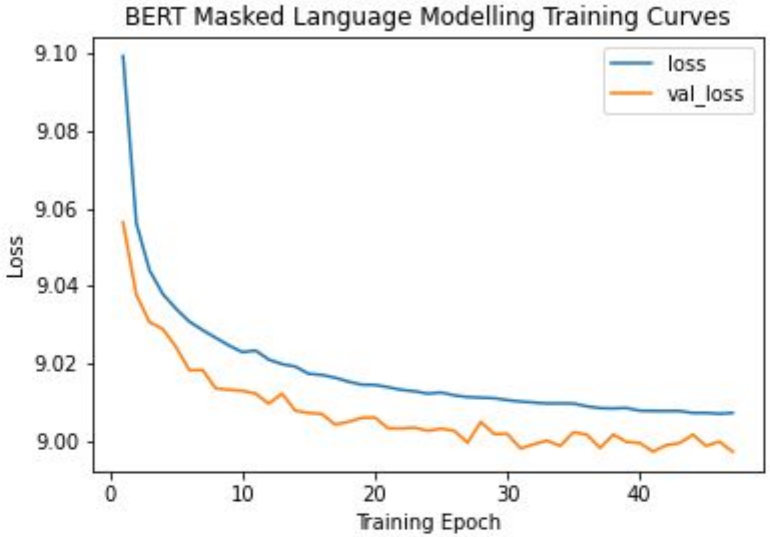


Attention Plot for GRU

Source https://www.tensorflow.org/tutorials/text/nmt_with_attention

BERT Masked LM





Model hyperparameters:

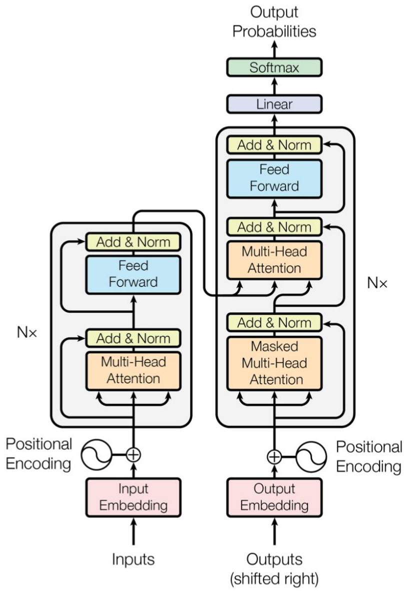
- **hidden_size: 128**
- **hidden_act: gelu**
- **hidden_dropout_prob: 0.1**
- **num_attention_heads: 2**
- **num_hidden_layers: 2**
- **Intermediate_size: 512**
- **Attention_probs_dropout_prob: 0.1**

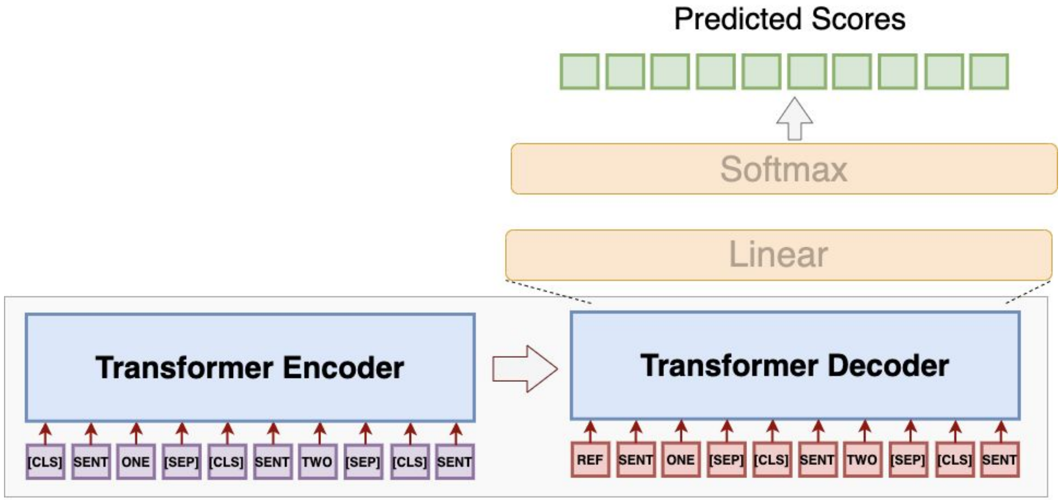
Source: https://arxiv.org/pdf/1810.04805.pdf

Transformer Encoder-Decoder

+ BERT Embeddings

+ Iterative Back-Translation



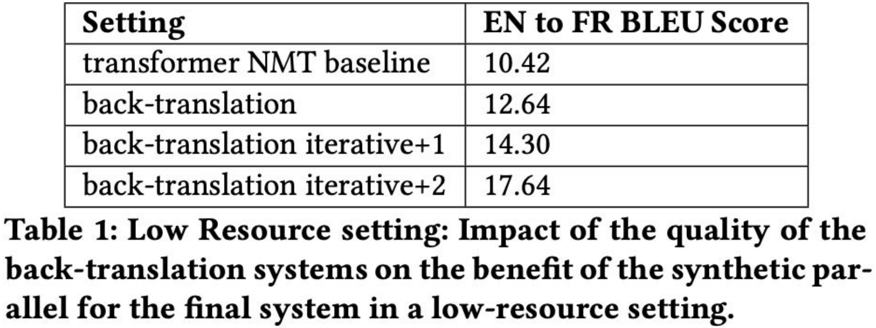


Transformer Model Architecture

Source: https://arxiv.org/abs/1706.03762

Transformer Encoder-Decoder

+ BERT Embeddings
+ Iterative Back-Translation



Setting	EN to FR BLEU Score
transformer NMT baseline	10.42
back-translation	12.64
back-translation iterative+1	14.30
back-translation iterative+2	17.64

Table 1: Low Resource setting: Impact of the quality of the back-translation systems on the benefit of the synthetic parallel for the final system in a low-resource setting.

Back-Translation data ratio, `1(real):6(synthetic)`

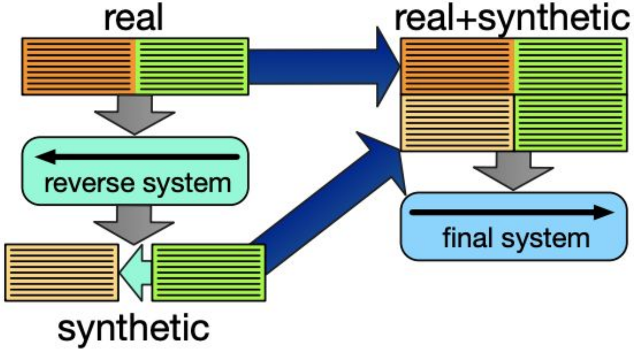
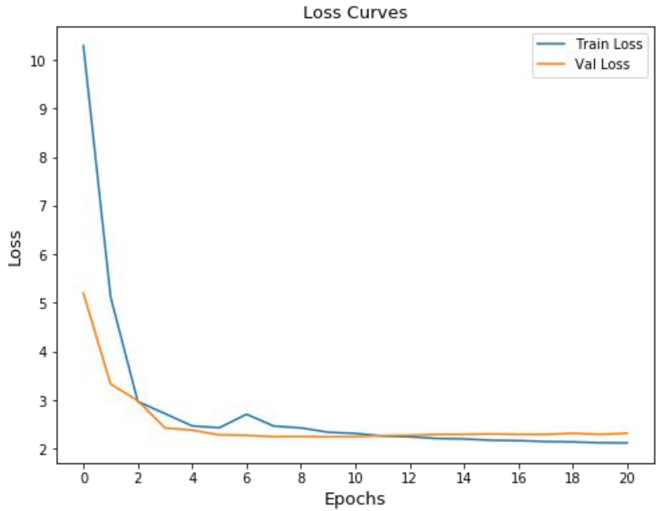


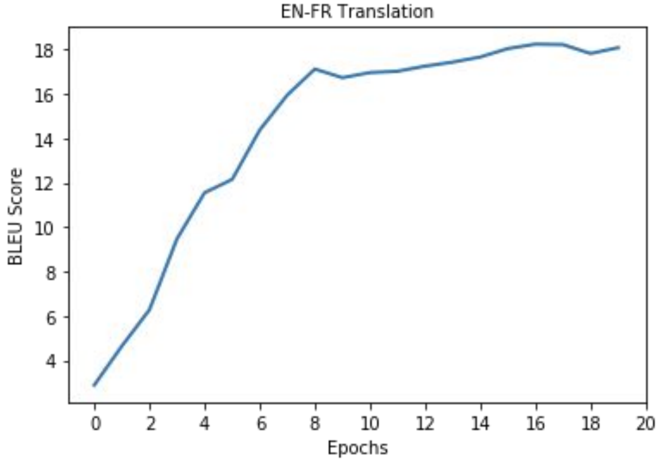
Figure 1: Creating a synthetic parallel corpus through back-translation. First, a system in the reverse direction is trained and then used to translate monolingual data from the target side backward into the source side, to be used in the final system.

Source: https://www.aclweb.org/anthology/W18-2703.pdf

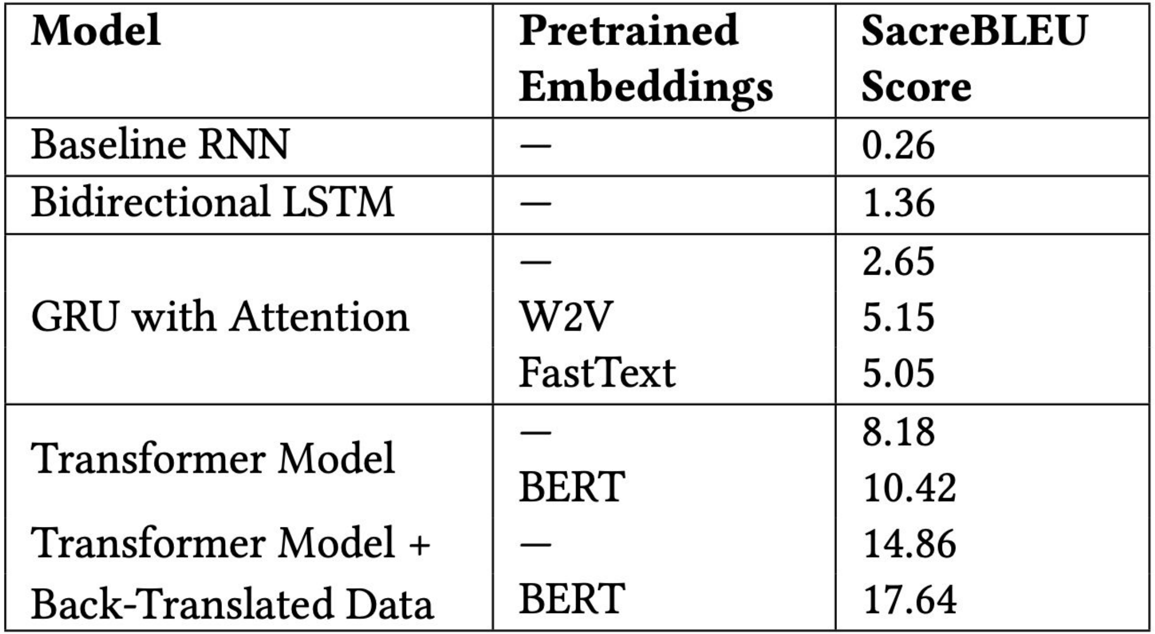
Model Hyperparameters

- Vocabulary size: 30k each for french and english
- Train-validation split, **`90:10`**
- Sequence length, **`seq_en=80, seq_fr=120`**
- Encoder & decoder - Identical layers stacked, **`N=4`**
- Outputs embedding dimension, **`d_model=128`**
- Number of parallel attention heads, **`A=8`**
- Position-wise feed-forward network, input and output dimension, **`d_model=128`**, and inner-layer dimensionality, **`d_ff=512`**
- Positional encodings - sine and cosine functions of dimension **`d_model=128`**
- Optimizer: Adam optimizer with **`lr=.001`**, **`β1=0.9`**, **`β2=0.98`** and **`ε=10`**
- Custom learning schedule as followed in *Vaswani et al. 2017*
- Residual dropout rate, **`P_dropout=0.1`**
- Back-translation data ratio**`, 1(real):6(synthetic)`**
- <u>Training on real+synthetic data</u>**`, Epochs=15`**
- <u>Final step of fine-tuning on true aligned data</u>**`, Epochs=5`**
- **`Batch_size = 128`**





Results

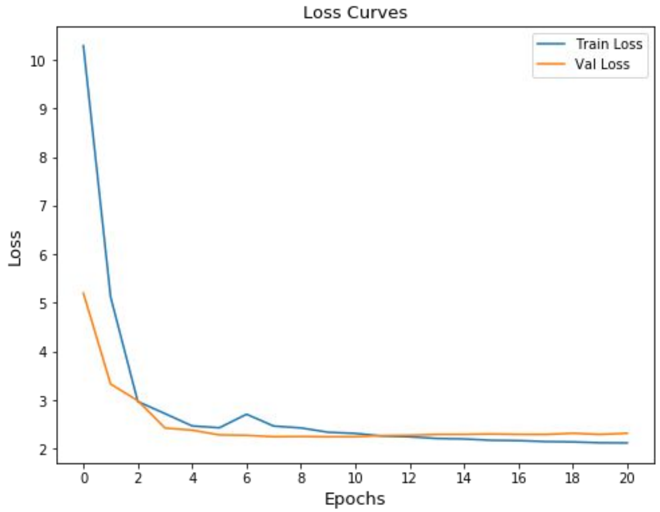


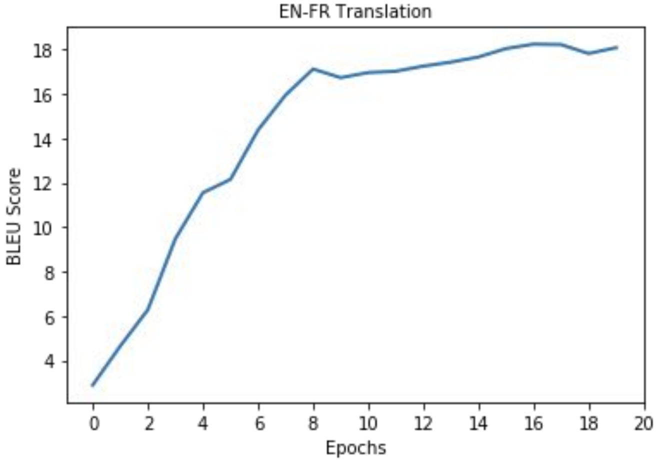
Model	Pretrained Embeddings	SacreBLEU Score
Baseline RNN	—	0.26
Bidirectional LSTM	—	1.36
GRU with Attention	—	2.65
	W2V	5.15
	FastText	5.05
Transformer Model	—	8.18
	BERT	10.42
Transformer Model + Back-Translated Data	—	14.86
	BERT	17.64

Table 3: BLEU Score summary of various models on held-out test set

Conclusion

- Transformer model outperformed all other models
- Best BLEU score of 17.64 on held-out validation set
- Iterative Back-translation greatly helped in low-resource setting
- Utilizing BERT Masked Language Modelling creates better representation of input tokens, improving performance
- Custom Tokenizer built over Byte-Pair Encoding helps us mitigate Out-Of-Vocabulary words problem, and better learn capitalization during target french generation
- More parallelizable and require significantly less time to train than RNNs and GRUs
- Larger amounts of synthetic data helped
- Best scores with back-translation ratio of `1(real):6(synthetic)`





Future Work

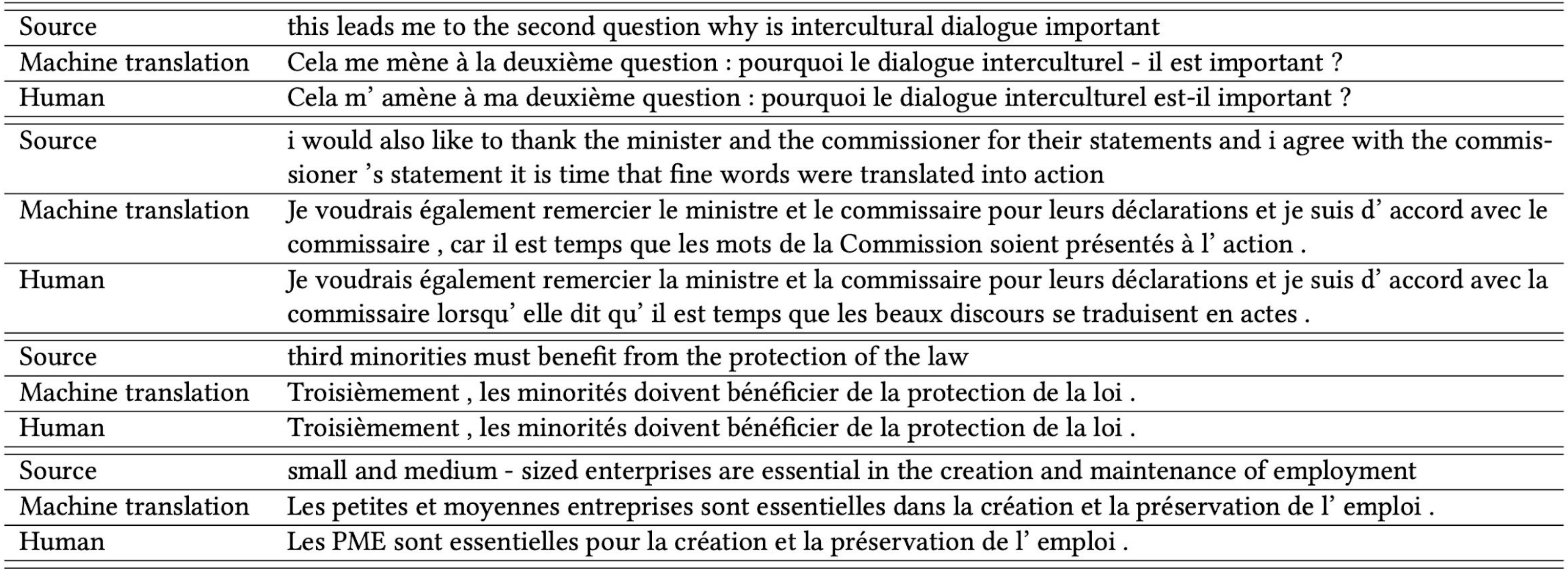
- Run more iteration of back-translation
- Experiment with higher capacity transformer model
- Utilize beam search along with length penalty to improve translation quality
- Employ label smoothing regularization technique. This hurts perplexity, as the model learns to be more unsure, but improves accuracy and BLEU score
- Deploy checkpoint averaging
- High capacity BERT MLM model, as the current model didn’t overfit on the current dataset
- Explore further multi-task hierarchical models, such as punctuation model, true-casing model, etc.
- Investigate multiple evaluation criterias to consider semantic meanings of generated translations
- Study Reinforcement Learning (RL) based approaches such as policy-gradient, REINFORCE etc. for low-resource NMT
- Examine model behaviour on small and large sentences independently

Q & A

Models Experimented

- RNN Encoder-Decoder
- Bidirectional LSTM
  - + ELMo
- GRU with Attention
  - + w2v
  - + fastext embeddings
- BERT Language Masking for pre-training
- Transformer Encoder-Decoder Model
  - + BERT pre-trained embeddings + Iterative Back-Translation

Sample Translations



Source	this leads me to the second question why is intercultural dialogue important
Machine translation	Cela me mène à la deuxième question : pourquoi le dialogue interculturel - il est important ?
Human	Cela m’ amène à ma deuxième question : pourquoi le dialogue interculturel est-il important ?
Source	i would also like to thank the minister and the commissioner for their statements and i agree with the commissioner ’s statement it is time that fine words were translated into action
Machine translation	Je voudrais également remercier le ministre et le commissaire pour leurs déclarations et je suis d’ accord avec le commissaire , car il est temps que les mots de la Commission soient présentés à l’ action .
Human	Je voudrais également remercier la ministre et la commissaire pour leurs déclarations et je suis d’ accord avec la commissaire lorsqu’ elle dit qu’ il est temps que les beaux discours se traduisent en actes .
Source	third minorities must benefit from the protection of the law
Machine translation	Troisièmement , les minorités doivent bénéficier de la protection de la loi .
Human	Troisièmement , les minorités doivent bénéficier de la protection de la loi .
Source	small and medium - sized enterprises are essential in the creation and maintenance of employment
Machine translation	Les petites et moyennes entreprises sont essentielles dans la création et la préservation de l’ emploi .
Human	Les PME sont essentielles pour la création et la préservation de l’ emploi .

Table 4: Sample translations of our best performing Transformer model.

Appendix

Transformer Attentions Plots

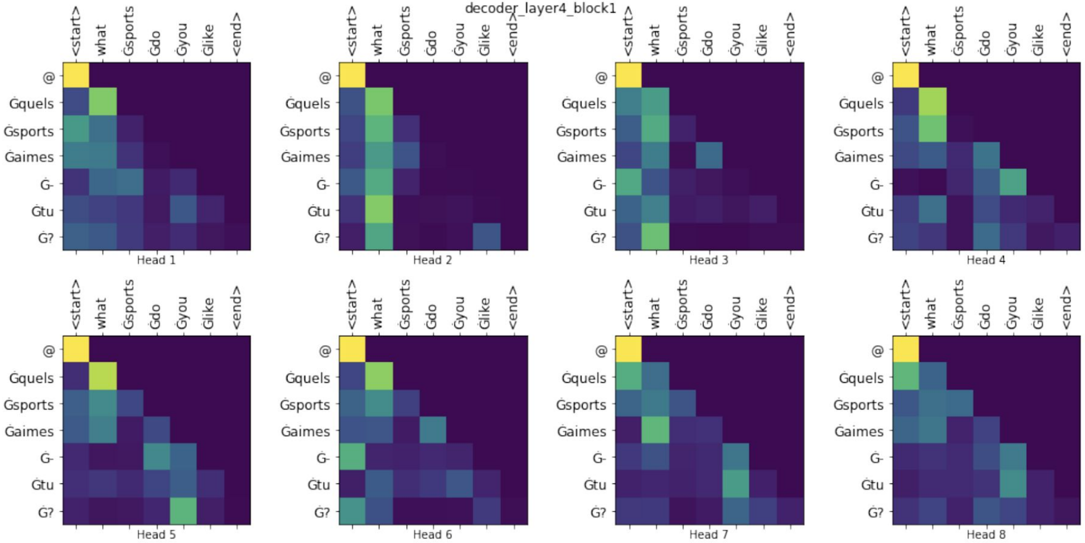


Figure 12: Masked Attention plot for decoder layer 4, block 1 of best performing transformer model