Low-Resources Machine Translation

IFT 6579 - Project 02 - Team 01

Chih-Chao . Carolyne . Akshay . Yifan . Ying

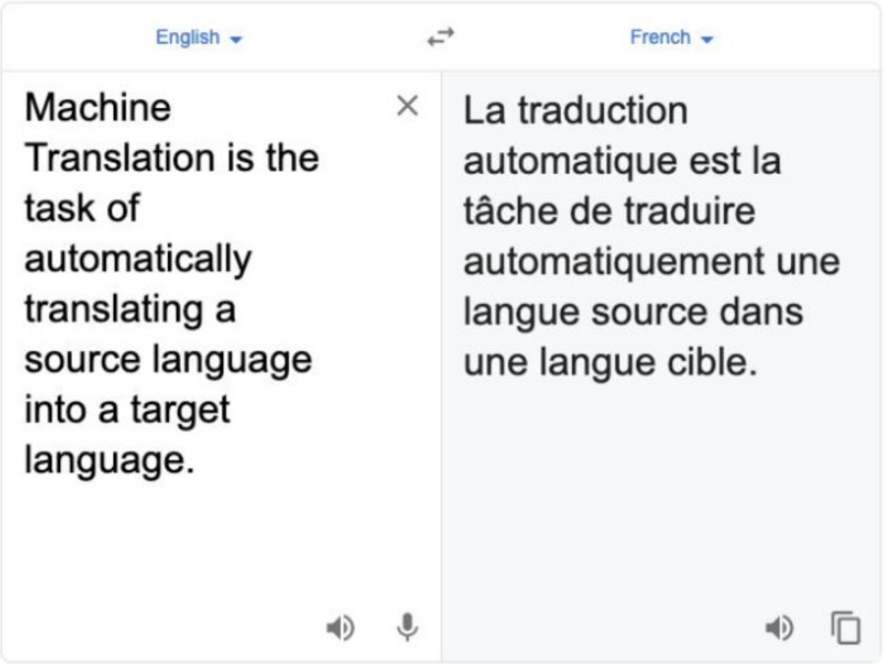
Introduction

- Traditional approaches(MLP) use phrase-based methods

- Our task: Build Neural Machine Translation (NMT) Create a French target sentence from an English source automatically. Take entire sequence to incorporate contextual information.

- Data organized in pairs of sentences for training, matched casing/punctuation:

  “Hello!” --> “Bonjour!”



Data Sources

1. 11K Aligned Parallel Examples:

   English mostly unpunctuated/lower-cased, French properly formatted. Both tokenized.

2. 474K Unaligned Monolingual Examples:

   Both properly formatted but untokenized.

Major Challenges

1. Low availability of aligned, parallel examples: Only 11k
2. Source English texts lacks formatting: uncapitalized, mostly unpunctuated
3. Massive amount of monolingual examples available but unaligned

Metric

- **B**i**L**ingual **E**valuation **U**nderstudy (BLEU) score

- sacreBLEU - a standardized implementation
  - Produces official WMT score
  - Standardizes tokenization handling
    - Computes BLEU
    - Outputs detokenized results
    - Downloads and manages of test sets.

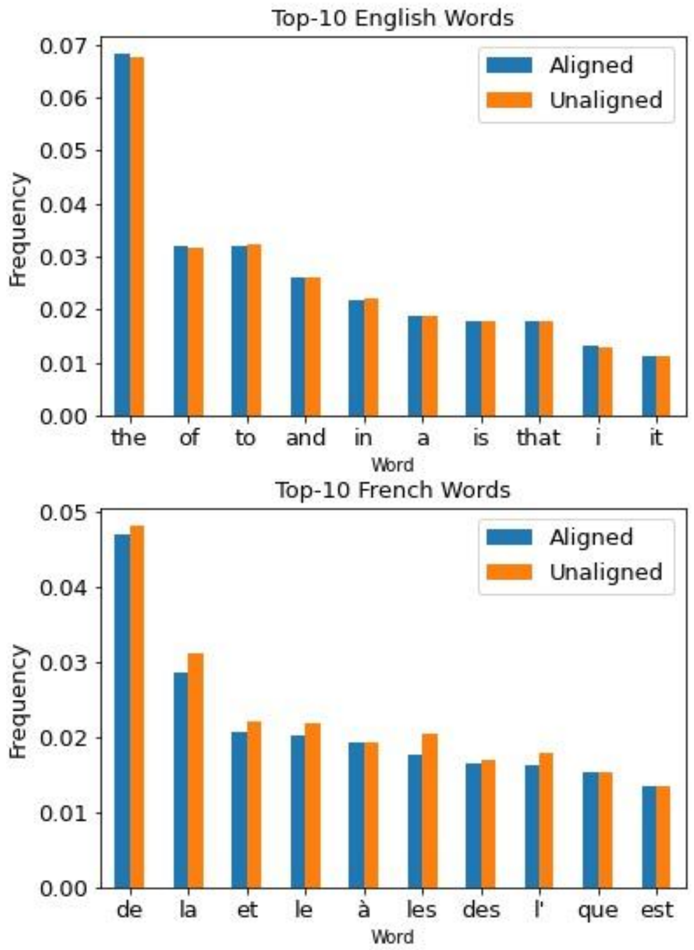
Data Analysis

- Word/Subword/Characters Tokens
    - Sentence lengths, vocab size, out-of-vocab(OOV): computational efforts - performance tradeoffs

- Exploratory Analysis:
    - Use given SpaCy tokenizer and punctuation remover
    - Word based unique tokens
    - OOV: non-alphabet words (~1.6-1.7% total word count (WC))

Data Analysis

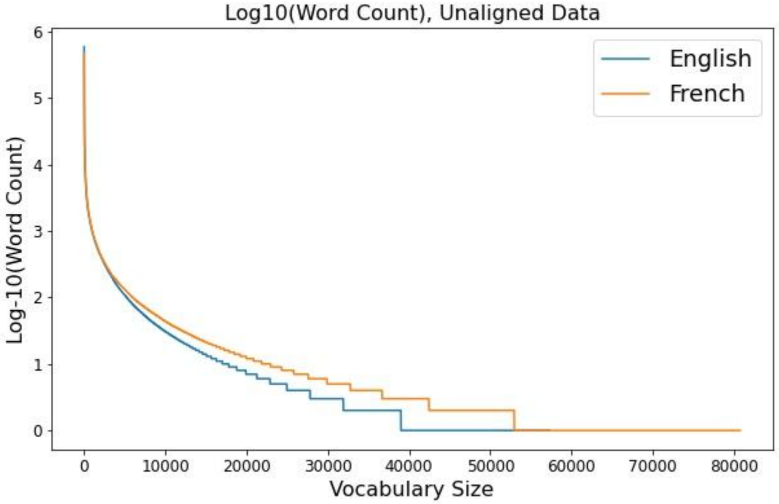
	WC	Vocab Size
Unaligned EN	8820344	57293
Aligned EN	205361	13657
Unaligned FR	9746232	80769
Aligned FR	227856	18222

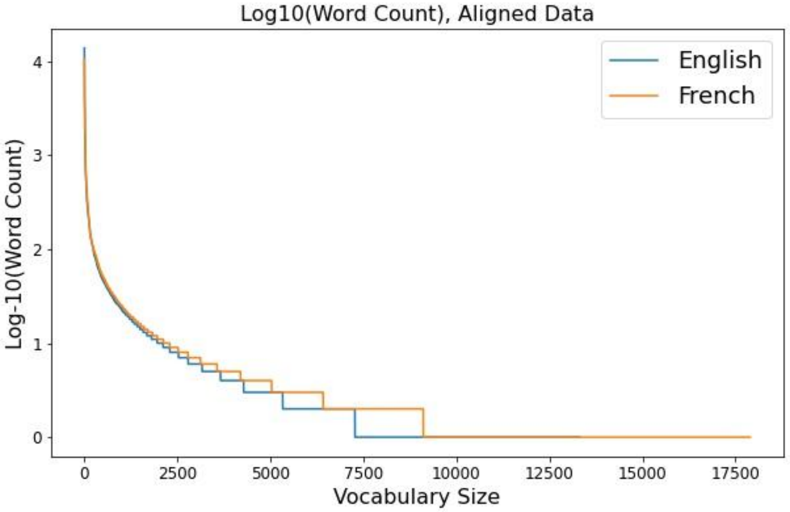


Data Analysis

Usable Vocab Size:

- Speed/Performance Tradeoff
- Log-10 transformation WC

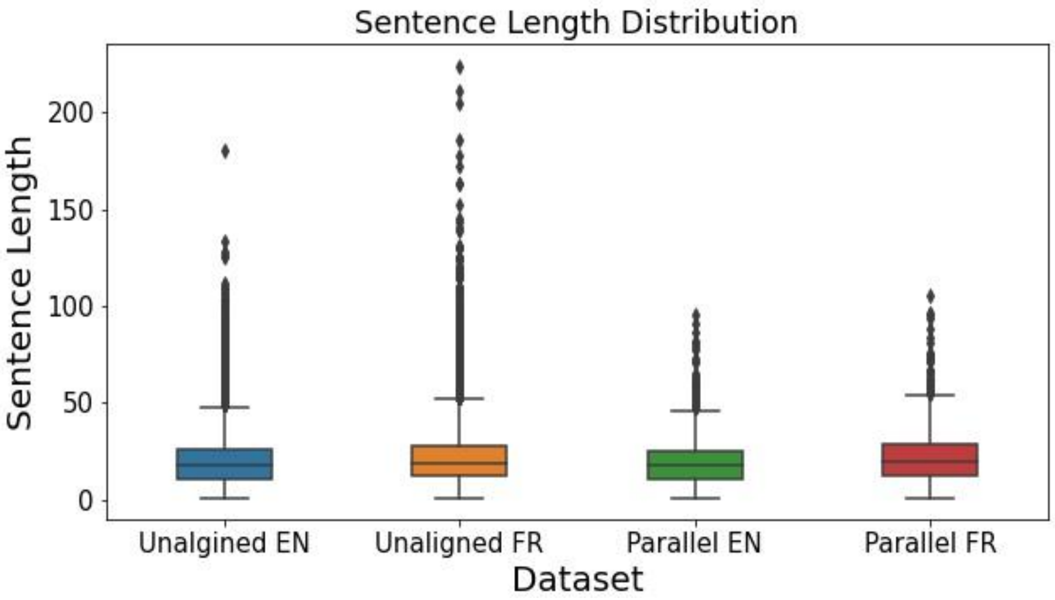




Data Analysis

Sentence Length:

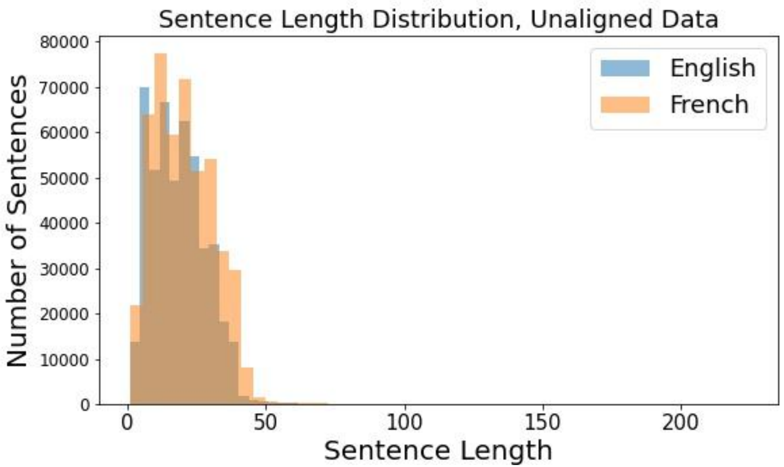
- Embedding sizes
- Language and dataset alignments

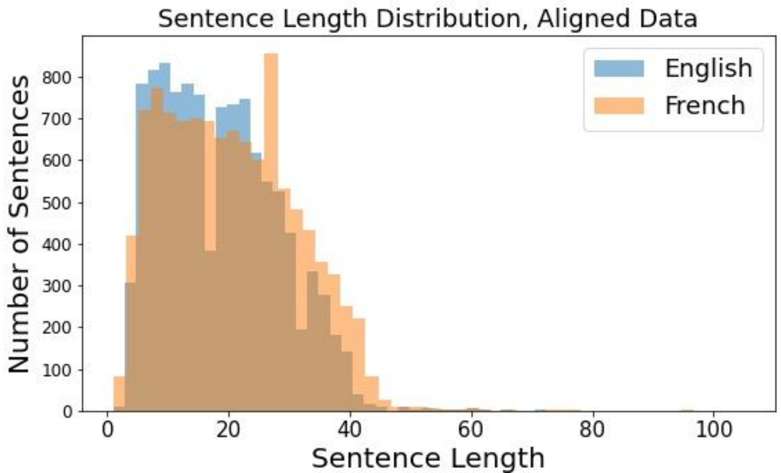


Data Analysis

FR/EN (mean/std/max):

- Unaligned: 20.56/10.82/224 | Aligned: 20.71/10.91/105
- Unaligned: 18.61/9.76/180 | Aligned: 18.61/9.76/95





Pipeline

- Data Pre-processing
- Word Embeddings
- Seq2Seq Model - (GRUs and Transformers)
- Back Translation

Data Pre-Processing

A vocab size of 20000 was selected for both English and French based on Data Analysis

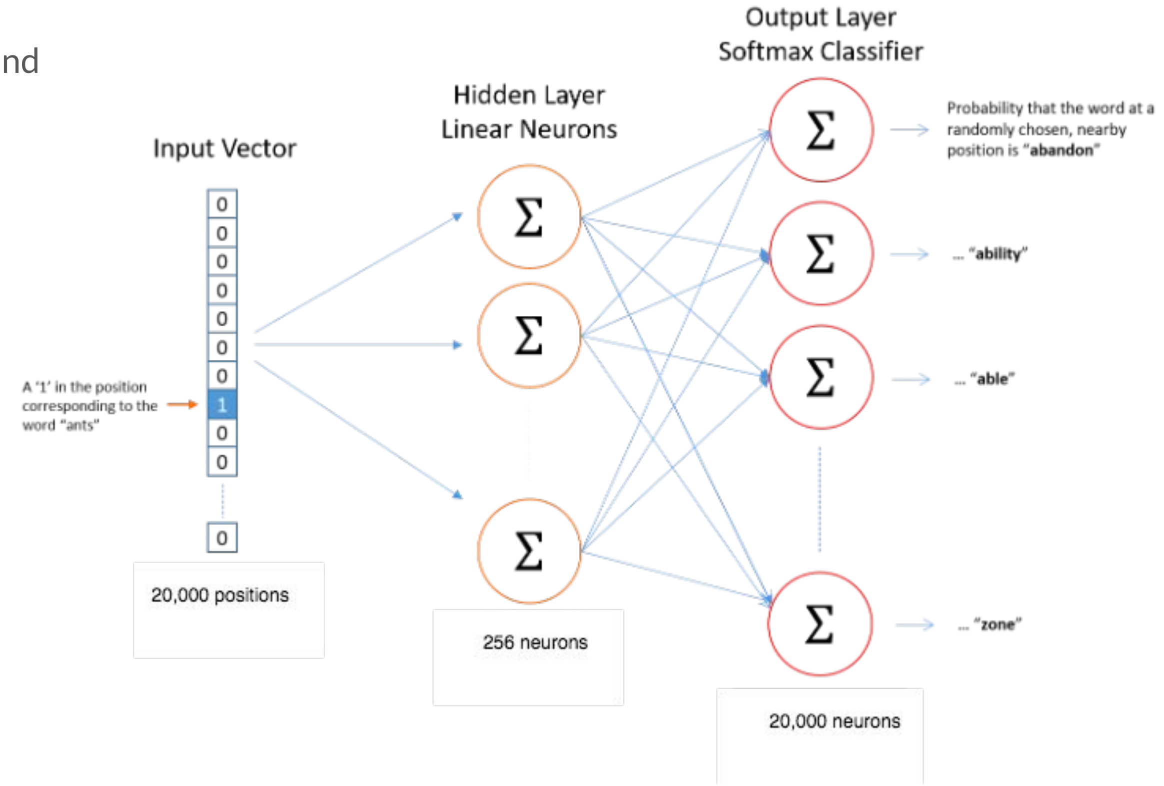
- For all data in both French and English
  - stripped each sentence of leading and trailing whitespace
  - added <start> and <end> tokens
- For the unaligned English data
  - lowercase the sentences
  - remove everything except for letters.
- For the unaligned French data
  - add space between the punctuation
- 80/20 split train/validation

Word Embeddings

Pre-trained **Word2Vec** on English and French unaligned text (CBOW)

Vocabulary size of 20K

Embedding Dimension: 256



Word Embeddings

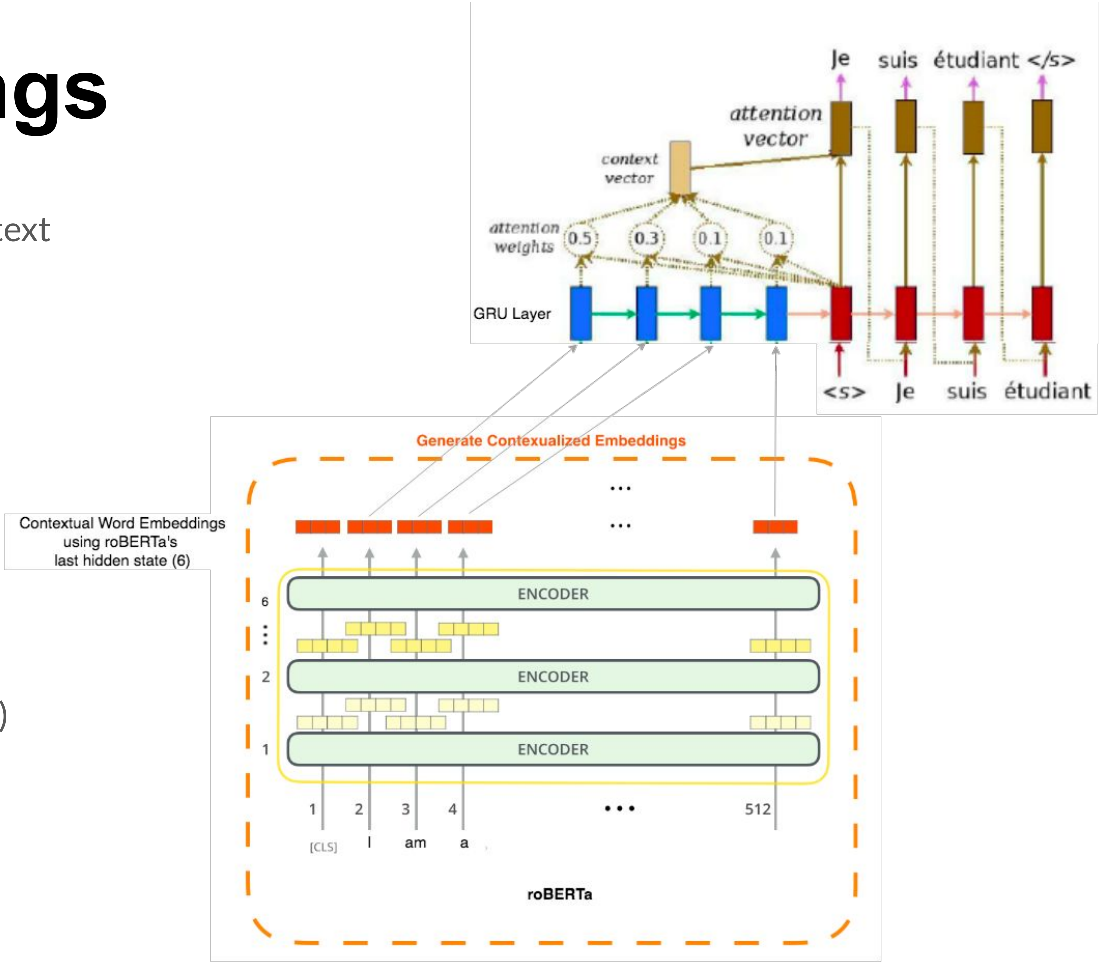
Pre-trained **roBERTa** on English unaligned text (masked language modeling task)

6 layers

512/128 max sequence length

52K/20K vocabulary size

10 epochs, batch size of 64 (16 for finetune)



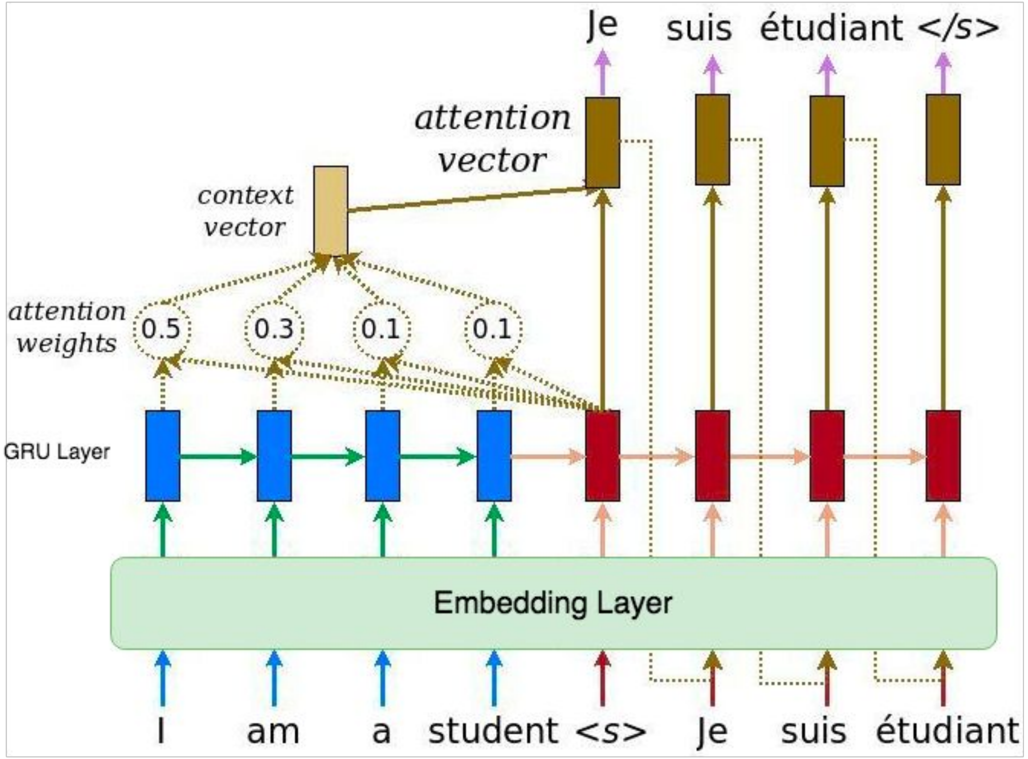
Seq-Seq GRUs with Attention

Adam optimizer with learning rate= 0.001, β_1= 0.9, β_2= 0.999, ε = 10^{-7}

Categorical Cross-Entropy loss between labels (gold French word) and predictions (French translated word from model)

512 hidden size

20 epochs, batch size of 64



Seq-Seq with Transformers

Adam optimizer with a custom learning r
schedule

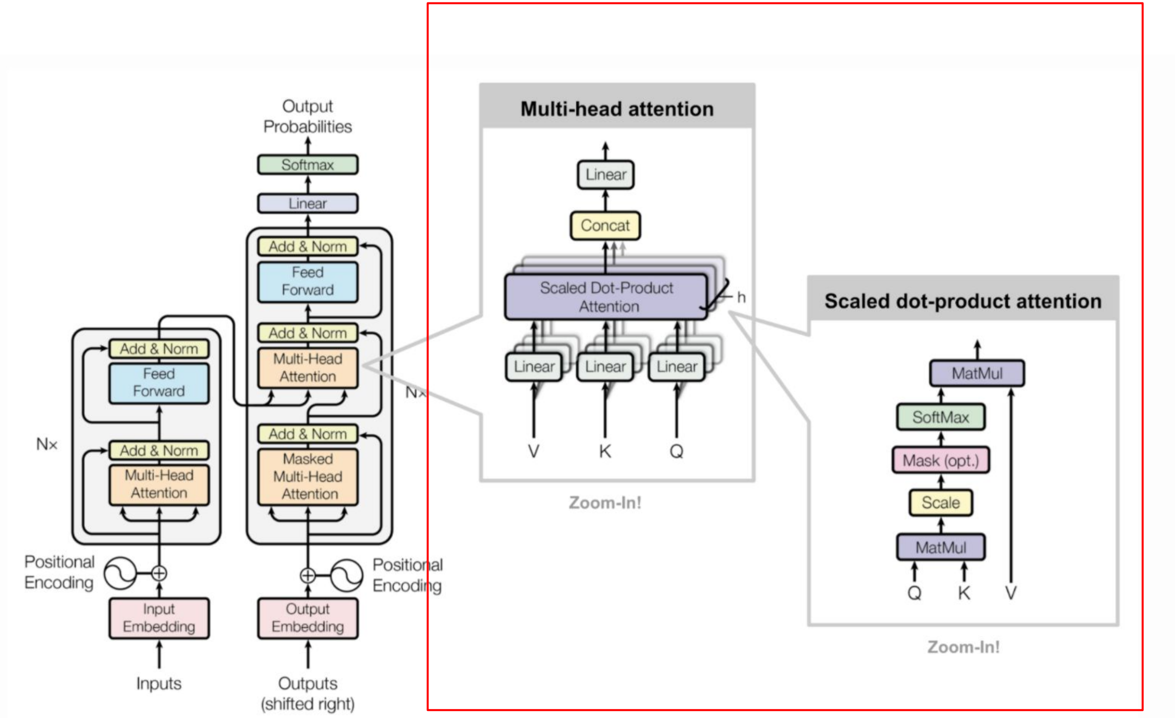
Categorical Cross-Entropy loss

1024 hidden size

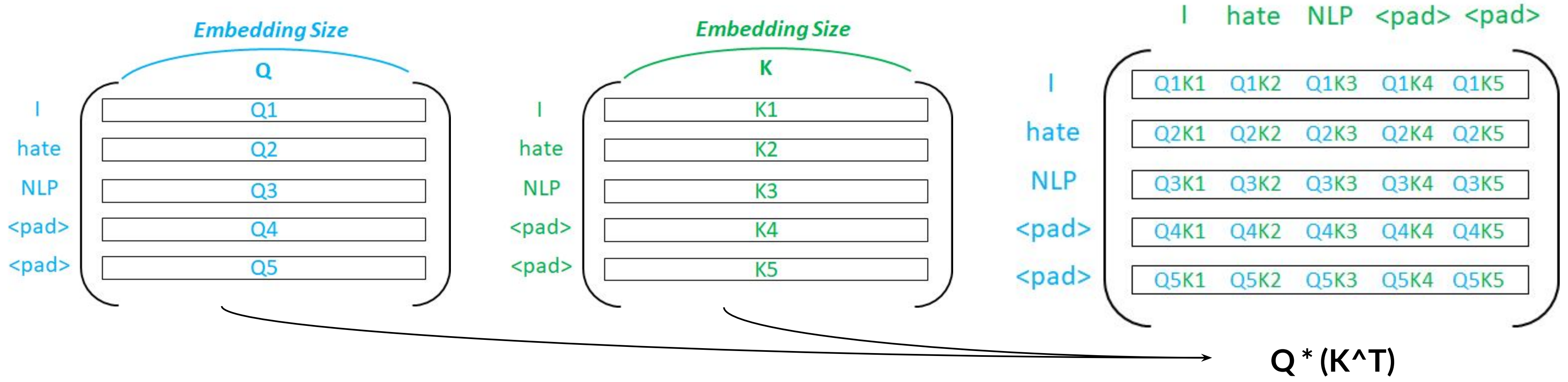
256 embedding dim

4 hidden layers for both encoder/decode

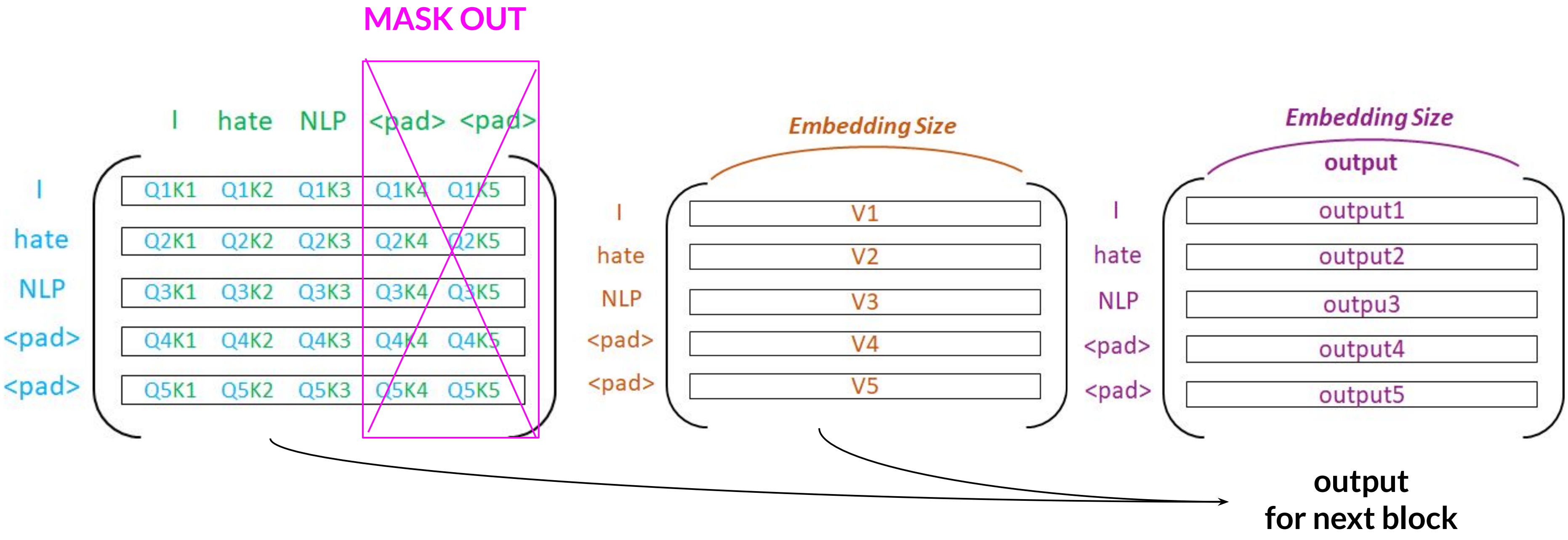
Batch size of 64



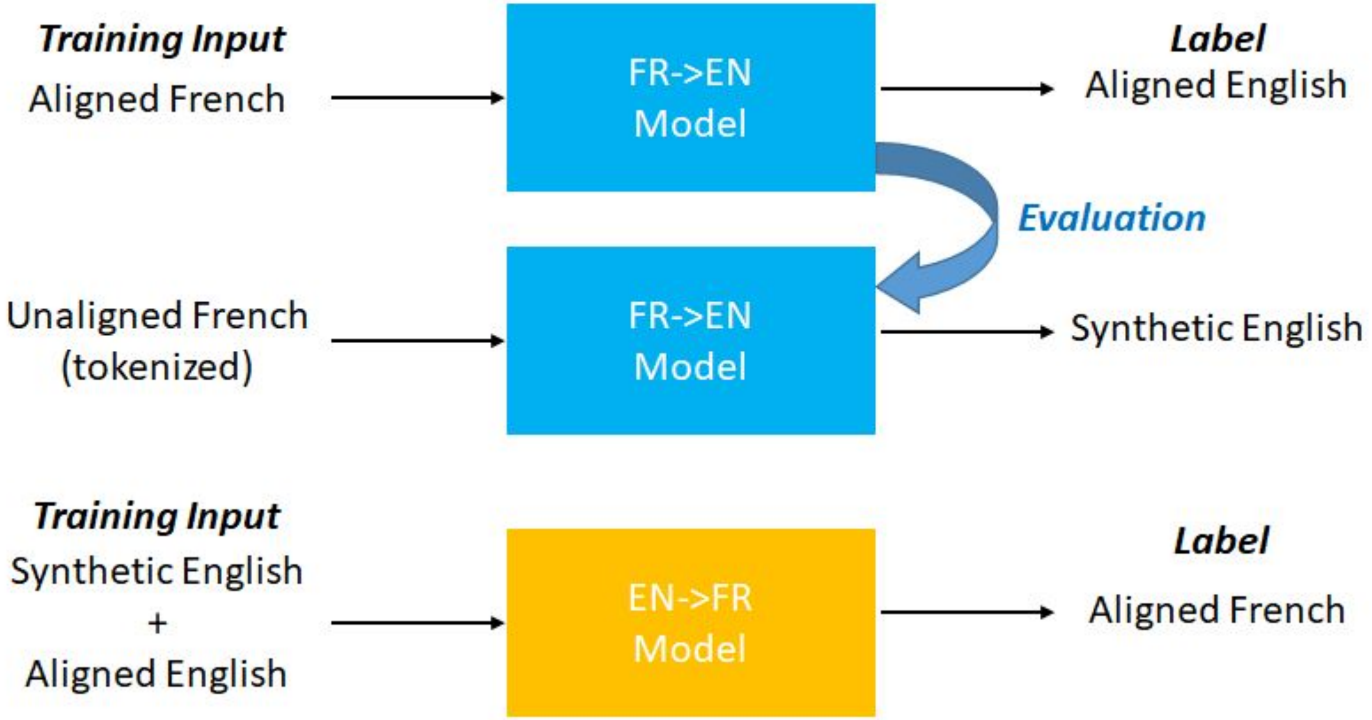
$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{n}}\right)\mathbf{V}$$



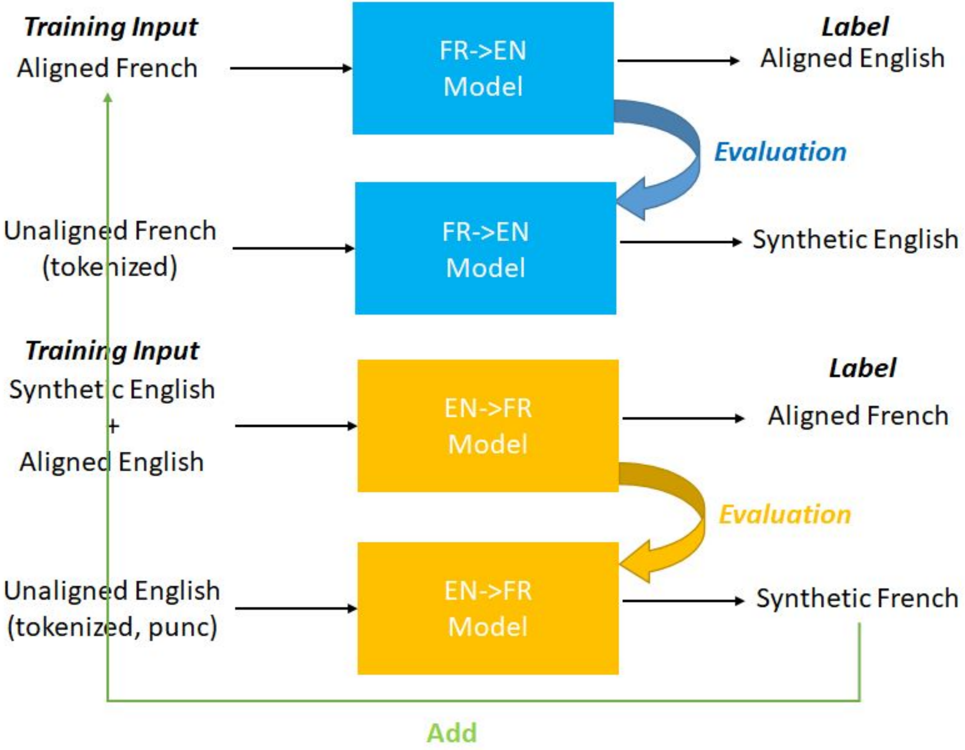
$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{n}}\right)\mathbf{V}$$



Back Translation



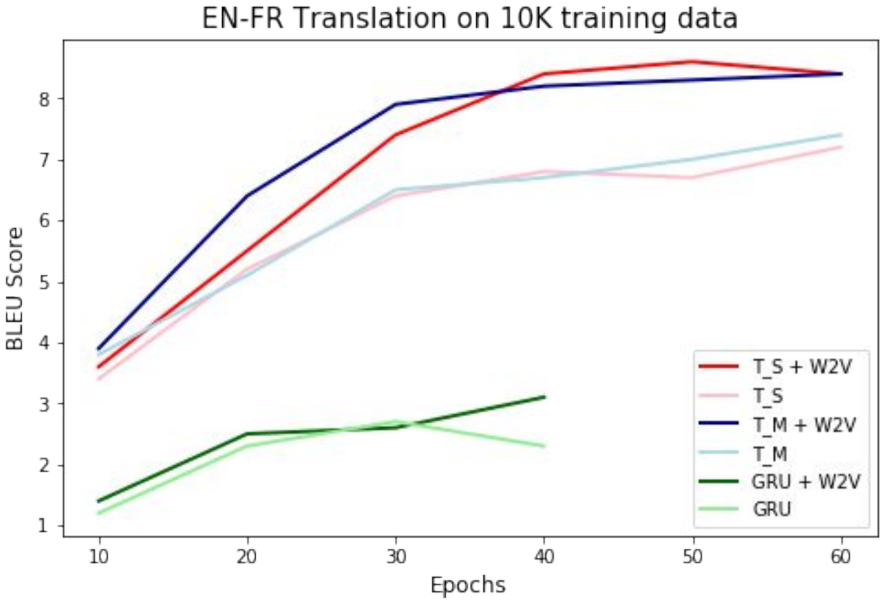
Iterative Back Translation



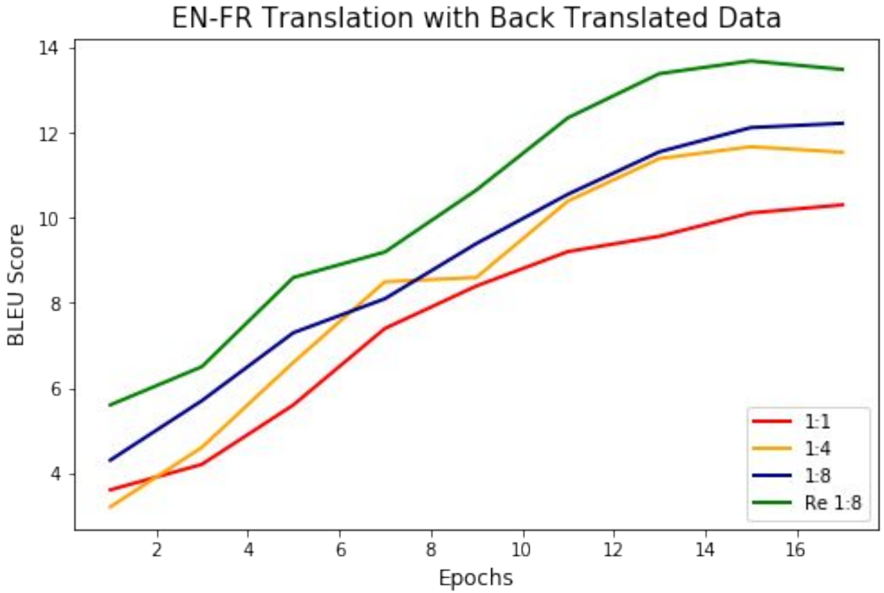
Results

ID	Models	Embedding-Size	Ratio	Bleu
01	GRU	Glorot-100	-	2.68
02	GRU	W2V-256	-	3.40
03	GRU	roBERTa_FE-768	-	0.72
04	GRU	roBERTa_FE-252	-	1.80
05	GRU	roBERTa_FT-252	-	2.40
06	T_S	-	-	7.20
07	T_M	-	-	7.48
08	T_S	W2V	-	8.38
09	T_M	W2V	-	8.40
10	T_M	roBERTa_FE	-	6.70
11	T_L	roBERTa_FE	-	3.32
12	T_S	W2V	BackTrans - 1:1	10.56
13	T_S	W2V	BackTrans - 1:4	11.47
13	T_S	W2V	BackTrans - 1:8	12.03
14	T_S	W2V	Re-BackTrans - 1:8	13.05
15	T_M	W2V	BackTrans - 1:1	10.64
16	T_M	W2V	BackTrans - 1:4	11.78
16	T_M	W2V	BackTrans - 1:8	12.13
17	T_M	W2V	Re-BackTrans - 1:8	**13.70**

Results

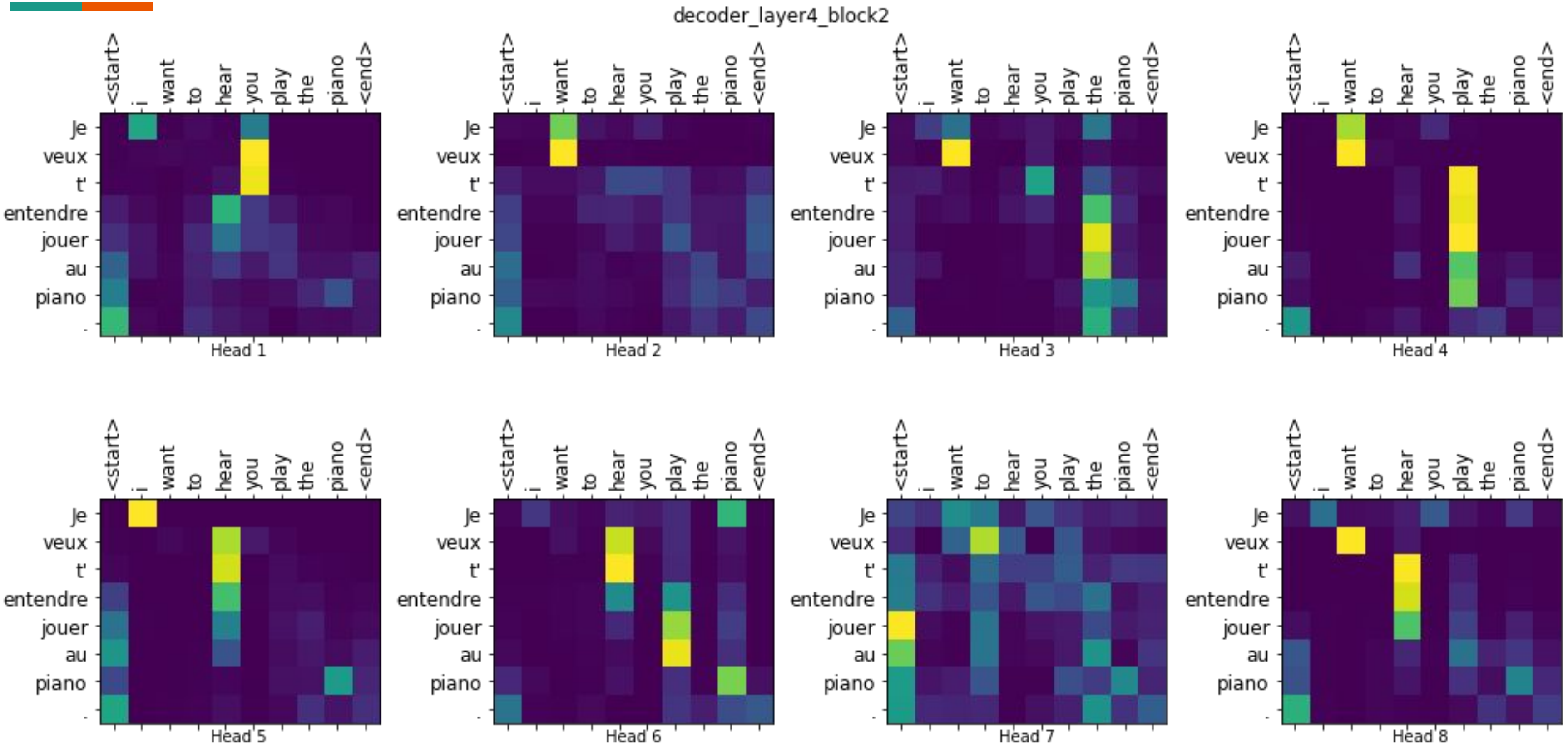


EN-FR Translation on 10K training data
BLEU Score
Epochs
T_S + W2V
T_S
T_M + W2V
T_M
GRU + W2V
GRU



EN-FR Translation with Back Translated Data
BLEU Score
Epochs
1:1
1:4
1:8
Re 1:8

Results



decoder_layer4_block2
<start>
i
want
to
hear
you
play
the
piano
<end>
Je
veux
t'
entendre
jouer
au
piano
.
Head 1
Head 2
Head 3
Head 4
Head 5
Head 6
Head 7
Head 8

Conclusion

- We investigated performance of a English to French NMT system using models with differing architectures like GRUs with attentions and transformer models.
- Low availability of aligned examples requires us to utilize unaligned data with large amount of back-translated synthetic data to increase training size, thus improving scores.
- Despite low quality, the model learns to construct sentences well and produced plausible results.
- We also showed while translation performance improves with additional synthetic data, performance tends to saturate when balanced is tipped too far in favour of synthetic data.
- Iterative back-translation was also explored where system improved with quality of synthetic data.

Future Work

- Beam Search decoder would be preferred over Greedy , and this is especially useful when we have synthetic samples in the training data.
- A multi-task system could be useful where BERT is trained on masked modelling and simultaneously improved by joining it with a decoder for translation task.
- We could also experiment with adding and fine tuning roBERTa in the decoder of GRU Seq2Seq model.
- We could explore various other tokenizers such as byte-processed ones to better handle OOV words.

Questions?