KUSHAGRA SETH
UCSC Id: 2005986
kuseth@ucsc.edu

Homework Assignment-1
(Spring 2023)
NLP-203

Professor: Ian Lane
TA: Geetanjali Rakshit

1 Part-1

Ipynb Notebook for Part-1: ./hw1_part1.ipynb

In the Tokenization step, the data is divided into train, dev, and test splits. This process is essential for ensuring that the models are trained, fine-tuned, and evaluated on different subsets of the data, preventing overfitting and promoting generalization.
To Preprocess the data, a Byte Pair Encoding (BPE) Tokenizer is used. This method helps in managing the vocabulary size and handling out-of-vocabulary (OOV) words, resulting in more efficient and accurate machine translation models.

1.1 Number of Tokens with BPE Tokenizer

Dataset	French	English
Train	3643031	3947880
Validation	24199	28498
Test	37629	40468

Table 1: Number of Tokens in diff fr-en data splits with BPE

1.2 Experiments with BPE Tokenizer

For this specific task, we trained two types of models:

1.) Convolutional Neural Network (CNN)
2.) Transformer encoder-decoder.

Both models were trained on the pre-processed data that was tokenized using the BPE Tokenizer. To evaluate the performance of these models, we used the SacreBLEU metric, which provides a standardized BLEU score for better comparison between different translation models.

Dataset	CNN	Transformer
Validation	28.99	29.04
Test	33.14	34.87

Table 2: SacreBLEU Scores with BPE

2 Part-2

Ipynb Notebook for Part-2: ./hw1_part2.ipynb

2.1 Number of Tokens with Moses Tokenizer

In the second part, the Moses tokenizer is utilized instead of the BPE tokenizer. The Moses tokenizer is a rule-based method that focuses on splitting text into words and punctuation marks. This approach is particularly effective for languages with clear word boundaries and relatively simple morphology.

Dataset	French	English
Train	3649842	3404336
Validation	21395	21149
Test	36025	33827

Table 3: Number of Tokens in diff fr-en data splits with Moses

BPE generates more tokens than Moses Tokenizer because it focuses on frequent character combinations, while Moses Tokenizer splits text into words. BPE can handle out-of-vocabulary words by breaking them into subword units, making it more robust for diverse language data. Moses Tokenizer, being rule-based, efficiently handles punctuation and capitalization but may struggle with unknown words. Choosing between the two depends on the task and the desired balance between model efficiency and language nuance capture.

2.2 Experiments with Moses Tokenizer

Similar to the first part, we trained both a Convolutional Neural Network (CNN) and a Transformer encoder-decoder for the translation task. Both models were trained on the preprocessed data, which was tokenized using the Moses tokenizer in this scenario. To evaluate the performance of these models and compare them with the BPE-tokenized counterparts, we again used the SacreBLEU metric, ensuring a standardized and consistent evaluation across all models.

Dataset	CNN	Transformer
Validation	24.08	27.51
Test	30.51	32.77

Table 4: SacreBLEU Scores w/o BPE

2.3 Examples from the Validation set

3 sentences that have better translations:

1.) And so I compared what I got to what I expected, and what I got was disappointing in comparison to what I expected.
2.) Not just bad : the worst book ever written .
3.) This picture reminded me of something .

3 Part-3

Ipynb Notebook for Part-3: ./hw1_part3.ipynb

3.1 Improvement using Shared Vocab

Shared source and target embeddings is used for improvement as a single embedding matrix for both the input (source) and output (target) languages in the encoder-decoder architecture. This approach

enables the model to learn a more compact and shared representation of the two languages, which can potentially lead to better generalization and translation quality.

3.2 Experiments: CNN vs Transformer

Dataset	CNN	Transformer
Validation	29.13	29.54
Test	33.74	35.13

Table 5: SacreBLEU Scores using Shared Vocab

In my experiments, I saw that using this technique improves the performance,i.e., the SacreBLEU score using Transformer to **35.13** using Transformer Encoder-Decoder Architecture.

3.3 Examples from the Validation set

3 sentences that have better translations:

1.) And I tried something interesting .
2.) There are some innovations in nuclear : modular , liquid .
3.) Now , what does that have to do with the placebo effect ?

1 sentence that has a worse translation:

1.) It 's like an apple on the ground .