Neural Machine Translation

Gujarati-English with transformer model

Oreen Yousuf & Jae Eun Hong

INDEX

- OVERVIEW
- EXPERIMENTAL SETUP
- RESULTS
- CONCLUSION

PROJECT OVERVIEW

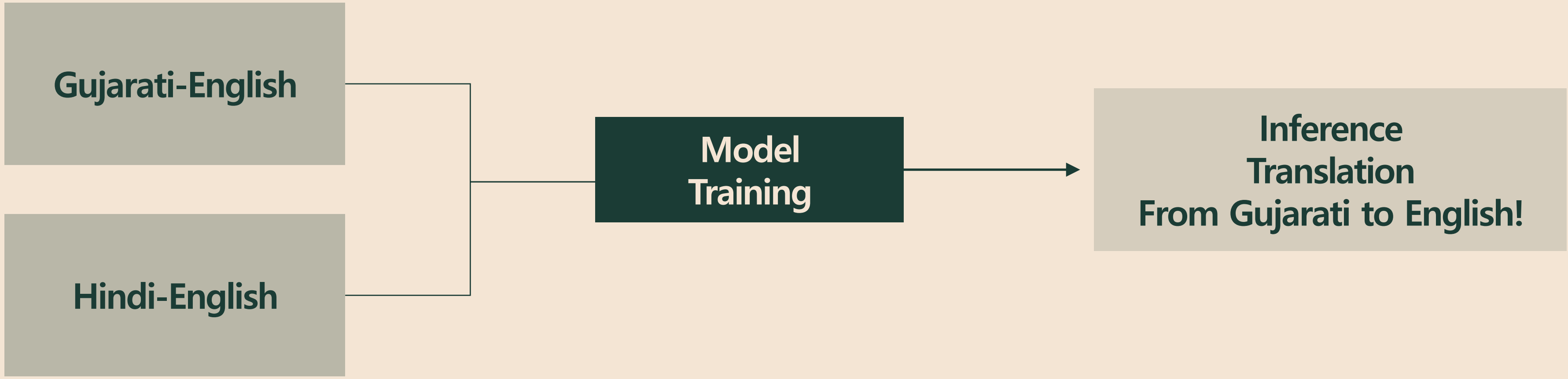
- Transformer-based Neural Machine Translation
- Gujarati – English parallel corpus
- Hindi – English parallel corpus
- Pre-trained models – baseline / **Multilingual models**
- Slightly increased BLEU score on the multilingual model
- Based on the paper ***<The IIIT-H Gujarati-English Machine Translation system for WMT19>***[1]

[1] *Goyal, Vikrant & Sharma, Dipti. (2019). The IIIT-H Gujarati-English Machine Translation System for WMT19. 191-195. 10.18653/v1/W19-5316.*

PROJECT OVERVIEW

Multilingual model

- Training a single model from multiple source languages into multiple target languages
- Exploiting data from other language pairs & joint training helps in improving the translation performance of NMT models



1) Goyal, Vikrant & Sharma, Dipti. (2019). The IIIT-H Gujarati-English Machine Translation System for WMT19. 191-195. 10.18653/v1/W19-5316.

EXPERIMENTAL SETUP

DATASETS – Train set

Gujarati –English corpus[1]

Dataset	Sentences
Wiki Titles v1.	11,671
Bible corpus	7,807
Clean crawled	10,650
Localization Opus	10,650
Wikipedia	18,033
Additional corpus[2]	65,000
Total	**123,811**

Hindi –English corpus

Dataset	Sentences
IIT Bombay Hi-En	1,609,682
Total	1,609,682

Total 1,830,480

↓

Total 1,421,612

2) http://www.statmt.org/wmt19/translation-task.html

3) https://github.com/shahparth123/eng_guj_parallel_corpus -by Language Processing Laboratory, Uka Tarsadia University, Gujarat

EXPERIMENTAL SETUP

DATASETS – Test set

Dataset	**Sentences**
News-test 2019	1,016
Total	1,016

EXPERIMENTAL SETUP

METHODS – Data Preprocessing

English

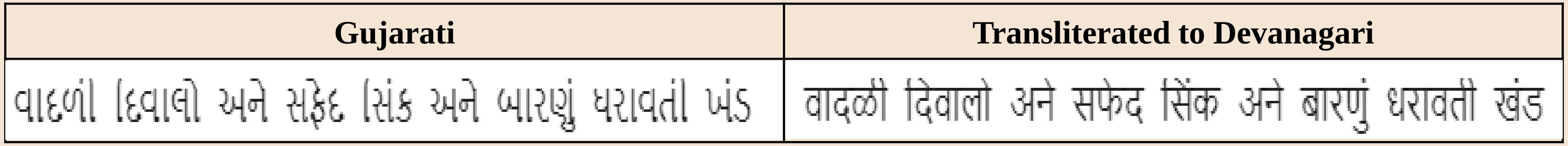
- **Moses Toolkit**

1. *Truecase*
2. *Tokenize*

Gujarati/Hindi

- **Indic NLP library**

1. *Normalize*
2. *Tokenize*
3. ***Transliterate***



Gujarati	Transliterated to Devanagari
વાદળી દિવાલો અને સફેદ સિંક અને બારણું ધરાવતી ખંડ	वादळी दिवालो अने सफेद सिंक अने बारणुं धरावती खंड

EXPERIMENTAL SETUP

METHODS – Data Preprocessing

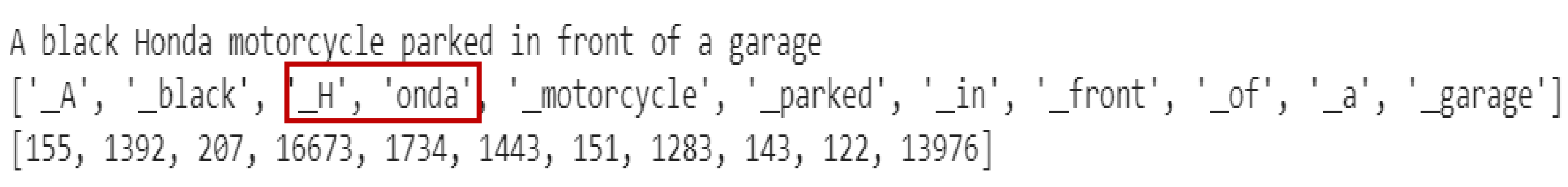
General Denoising

Deleted data	Sentence size after deletion
NULL data	1,725,192
Sent pair w/ foreign char ratio	1,723,259
Sent length > 120	1,421,612
Final train data size	**1,412,612**

EXPERIMENTAL SETUP

METHODS – Sub-word tokenization w/ Byte-Pair Encoding

	Baseline	Multilingual
Vocab size	32000	
Vocabularies from	Hindi/English	Gujarati+Hindi/English



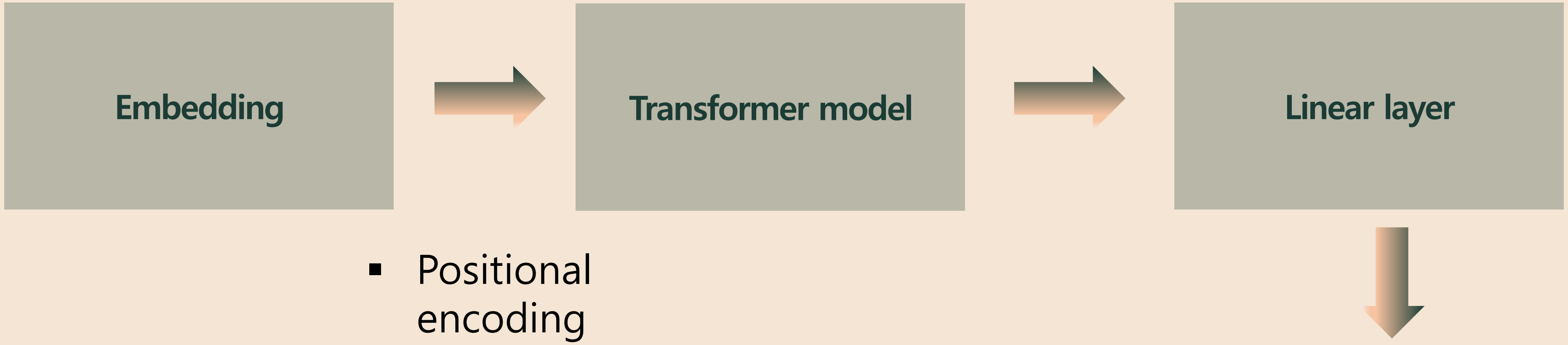
```
A black Honda motorcycle parked in front of a garage
['_A', '_black', '_H', 'onda', '_motorcycle', '_parked', '_in', '_front', '_of', '_a', '_garage']
[155, 1392, 207, 16673, 1734, 1443, 151, 1283, 143, 122, 13976]
```

EXPERIMENTAL SETUP

METHODS – Train architecture in Pytorch[3]

Model

Seq2Seq network using Transformer



- Positional encoding

- Unnormalized probabilities for each token in the target lang

4) https://pytorch.org/tutorials/beginner/translation_transformer.html

EXPERIMENTAL SETUP

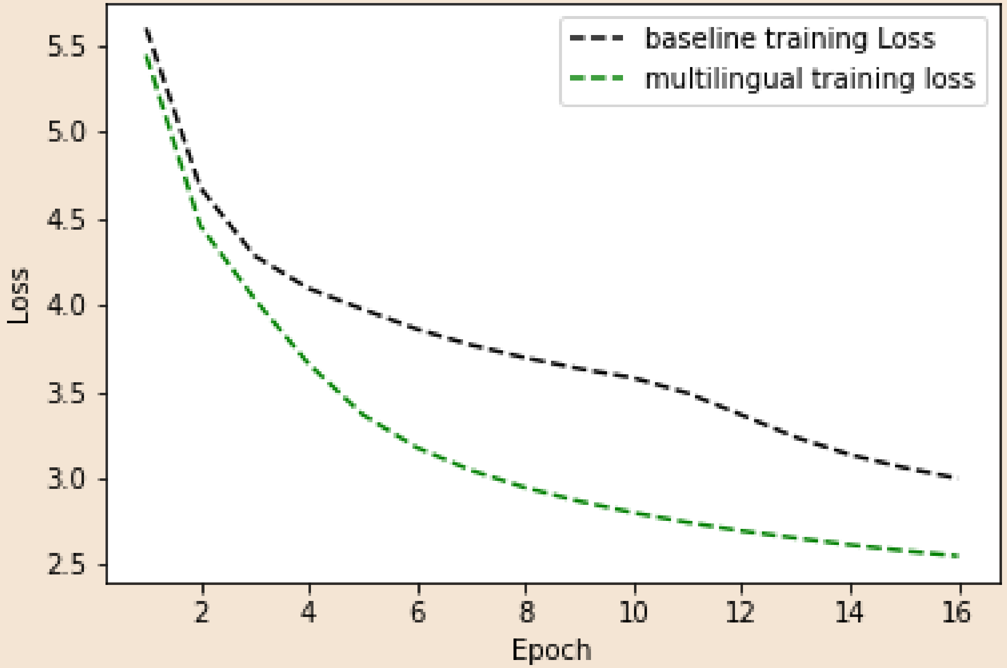
METHODS – Train architecture in Pytorch[3]

Hyperparameters

	Baseline	Multilingual
Source vocab size	32000	
Target vocab size	32000	
Embedding size	512	
Number of Heads	8	
Feed forward dim	512	
Batch size	64	
Encoder layers	6	
Decoder layers	6	
Epoch	16	
Optimizer	Adam	

RESULTS

Training loss



RESULTS

BLEU on Test set

	Baseline	**Multilingual**
BLEU score	0.03494	**0.04453**

CONCLUSION

SHORTCOMINGS & FUTURE WORK

Data augmentation

Transfer learning

Pivot-language on MNMT model

Hyper-parameter optimization

Fine-tuning

Q & A