Machine Back-translation using NMT

[Group 9]

Nisarg Doshi	202111033
Divya Shah	202111040
Harsh Prajapati	202111074

Nov 20, 2022

Introduction and Motivation

Introduction:

- With advent of deep learning, neural machine translation (NMT) performs better than statistical machine translation(SMT).
- For training neural machine translation models, we need high quality parallel data.

Motivation:

- Machine translation is a difficult task for some languages as they are resource-poor.
- Researchers[1] have shown, backtranslating monolingual data and combining it with authentic parallel data can train high quality NMT systems.

[1] Poncelas, Alberto et al. "Investigating Backtranslation in Neural Machine Translation." *ArXiv* abs/1804.06189 (2018)

Problem Statement

- To investigate how using back-translated data as a training corpus, both as a separate standalone dataset as well as combined with human-generated parallel data, affects the performance of an NMT model.

- Also, to figure out unknown factors regarding the actual effects of back-translated data on the translation capabilities of an NMT model for different languages.

Project Workflow

EN(base) -> DE(translated)

DE (base) -> EN(translated)

DE(translated) -> EN (synthesized)

EN(base) + EN (synthesized) -> DE (final)

Dataset and Model

Dataset:

- WMT Training Dataset[1]: English-German dataset
- Contains 4.48M parallel English-German sentences

Preprocessing:

- Dataset is tokenized, truecased and shuffled.

Model and tools:

- Tool: OpenNMT library[2]
- Model: Bi-LSTM with 500 hidden units and vocabulary size of 50k.

[2] Bojar, Ondřej, et al. "Findings of the 2014 workshop on statistical machine translation." *Proceedings of the ninth workshop on statistical machine translation*. 2014.
[3] Klein, Guillaume, et al. "Opennmt: Open-source toolkit for neural machine translation." *arXiv preprint arXiv:1701.02810* (2017).

Continue - Dataset and Model

Epoch :

- German to English : 90000 Epochs
- English to German : 90000 Epochs

Model Accuracy :

English to German :

- Training Accuracy : 47.96
- Validation Accuracy : 50.56

German to English :

- Training Accuracy : 49.98
- Validation Accuracy : 52.53

Methodology

Authentic data only:

- Trained only with authentic data available
- Used as benchmark

Synthetic data only

- When no parallel data is available.
- Used as reference for worst-case scenario for quality of data.
- Can be used for resource-poor languages.

Hybrid data:

- Combination of authentic and synthetic data
- Goal is to analyze results for different proportion of authentic to synthetic data

Methodology

Set-up

- Followed default training configuration of OpenNMT guidelines.

Training

- English to German model
- German to English model

Testing

- All models are evaluated using 3 different test files from news domain.

Validation

- Models are validated using news test set 2015

Evaluation Metric

- BLEU Score

Testing Dataset

Test Dataset (News)	**Number of sentences**
News Test Set- 2012	3003
News Test Set- 2013	3000
News Test Set-2014	2737

Result

Test Dataset (News)	**Model**	**English to German BLEU score**	**German to English BLEU score**
News Test Set-2012	BI-LSTM	10.35	9.46
News Test Set-2013	BI-LSTM	8.56	5.51
News Test Set-2014	BI-LSTM	11.66	15.08

Future Work

- Evaluation of hybrid corpus with different proportion of synthetic (back-translated) data.

- Repeating this experiment with Transformer.

- Repeating this experiment for Indian languages.

References

[1] Poncelas, Alberto et al. “Investigating Backtranslation in Neural Machine Translation.” *ArXiv* abs/1804.06189 (2018)

[2] Bojar, Ondřej, et al. "Findings of the 2014 workshop on statistical machine translation." *Proceedings of the ninth workshop on statistical machine translation*. 2014.

[3] Klein, Guillaume, et al. "Opennmt: Open-source toolkit for neural machine translation." *arXiv preprint arXiv:1701.02810* (2017).

Thank You

BLEU Score(Evaluation Metric)

- The measure BLEU (BiLingual Evaluation Understudy) is used to evaluate machine-translated text automatically.
- It gives value in number between 0 and 1.
- A value of 0 indicates low quality translation i.e. no overlap between machine translation and reference translation.
- A value of 1 indicates high quality translation i.e. perfect overlap between machine translation and reference translation.