NLP Capstone Project

Model Cards

Olatomiwa Akinlaja, Kudzai Sibanda

Witwatersrand University

Model Card 1:JoeyNMT

Model details :

- baseline code provided by MASAKHANE; a research effort for NLP for African languages
- The transformer-based encoder-decoder model was introduced by Vaswani et al. implemented with JoeyNMT; Initially developed by Jasmijn Bastings & Julia Kreutzer, now maintained by Mayumi Ohta [1]

Intended use:

- Educational purposes and NLP research
- Particularly intended for African language translation

Factors:

- It is often challenging to implement and modify classic NLP architectures.
- Can be greatly affected by poor tokenization

Metrics:

- BLEU Score [2]

Training & Evaluation Data:

- Umsuka English - isiZulu Parallel Corpus

Ethical Considerations:

- all code and information used is publically available so there are no ethical violations.

Caveats and recommendations

- Notebook intended to allow the use of custom parallel data.
- Struggled at forming a proper corpus
- Use larger dataset

Model Card 2:Huggingface

Model details :

- High performance on NLU and NLG tasks
- Low barrier to entry for educators and practitioners
- Transformer [3]

Intended use:

- State-of-the-art NLP for everyone

Factors:

- Training a machine translation model is usually time consuming

Metrics:

- BLEU Score [2]

Training & Evaluation Data:

- Umsuka English - isiZulu Parallel Corpus

Ethical Considerations:

- Data and other information freely available to the public. No ethical violations noted.

Caveats and recommendations

- The pre-trained m2m100 state-of-the-art English-isiZulu translations without the need of an english training corpus
- Use different pre-trained models

Model Card 3: Seq2Seq

Model details :

- Personal attempt at creating a seq2seq encoder-decoder model
- LSTM neural machine translation using keras

Intended use:

- Neural Machine translation.

Factors:

- Did not use Attention in the model

Metrics:

- BLEU Score [2]

Training & Evaluation Data:

- Umsuka English - isiZulu Parallel Corpus

Ethical Considerations:

- None encountered, all data is freely available online.

Caveats and recommendations:

- The pre-trained m2m100 state-of-the-art English-isiZulu translations without the need of an English train corpus
- Use attention mechanism