UNIVERSITY OF PRETORIA

Weekly Team Updates

TEAM: Augmentation

12 JUN'24

Weekly Goals for the Team -

MEMBERS:

- JOSHUA
- SRINIVAS
- SANJAY
- ISHAAN

1. REVIEW PAST AUGMENTATION LITERATURE
2. BRING TOGETHER DATASETS AND AUGMENTATION TECHNIQUE TO PROPOSE NEW TECHNIQUES
3. PRESENT THE WORK IN CONCISE MANNER

TIMELINE : JOSHUA

CoSDA-ML

Multi-Lingual Code-Switching Data Augmentation for Zero-Shot Cross-Lingual NLP

- **Requirements**:
- Source Data
- Target Data
- Source->Target Dictionary
- Dataset Utility Tool

- **Source Data, Target Data:** DeepKIN
- A deep learning toolkit for Kinyarwanda NLP.
- They use a Google-translated version of the GLUE benchmark tasks (MRPC, RTE, STS-B, SST-2, QNLI) as well as Tweet Sentiment Analysis to fine tune KinyaBERT.

Timeline : Joshua

CoSDA-ML

Multi-Lingual Code-Switching Data Augmentation for Zero-Shot Cross-Lingual NLP

- **Dictionary:** PanLex
- **Source**: English (10280)
- **Target**: Kinyarwanda (341)

- **Model**: Multi-Class Sentiment Classification (+ve, -ve, neutral)
- SC4 mBERT model trained on the custom Kinyarwanda Dataset

TIMELINE : JOSHUA

NOVELTY

Low Resource Augmentation

Traditional cross-lingual NLP techniques often require parallel corpora for training, which are scarce for Kinyarwanda. CoSDA-ML, through its innovative use of code-switching data augmentation, eliminates the need for such resources. Instead, it relies on dictionary-based augmentation, making it feasible to enhance Kinyarwanda's NLP capabilities without extensive bilingual datasets.

Improved Sentiment Analysis

Sentiment analysis in Kinyarwanda is challenging due to the lack of annotated sentiment datasets. By using the CoSDA-ML framework, the sentiment analysis model can be fine-tuned using code-switched data. This approach improves the model's ability to understand and process sentiment in Kinyarwanda, even when trained primarily on data from other languages.

Timeline : Joshua

REFERENCES

CoSDA-ML

Paper: https://arxiv.org/pdf/2006.06402
Git: https://github.com/kodenii/CoSDA-ML?tab=readme-ov-file

DeepKIN

Paper: https://arxiv.org/abs/2203.08459
Git: https://github.com/anzeyimana/DeepKIN/tree/main

PanLex

Paper: https://aclanthology.org/I17-1037/
Git: https://github.com/dylandilu/Panlex-Lexicon-Extractor

Timeline : Srinivas

1

Using the Serengeti-E250 model for data augmentation

2

Random sentences were chosen, and certain words were replaced with mask tokens

3

After augmentation, labels from the original sentence re-used and added to the original dataset

4

Original dataset contains 3302 entries, augmented by 20% - final dataset contains 3962 entries

Timeline : Sanjay

UXLA(Unsupervised cross lingual augmentation):

The paper aims to solve a problem called zero-resource cross-lingual transfer, which means adapting a model trained on one language to perform a task on another language without using any labeled data in the target language.

Methodology- The paper uses a multilingual masked language model (XLM-R) to generate new sentences in both the source and target languages. The paper uses two techniques to improve the model's performance on the target language: data augmentation and self-training.
The paper uses two techniques to select the most reliable examples from the unlabeled data: co-distillation and co-guessing.

Shortcomings-
1)The paper does not consider the effect of different pretraining objectives or architectures of the multilingual masked language model. This means that the results might vary if a different model was used. 2)The paper does not compare UXLA with other unsupervised methods for cross-lingual adaptation.