University of Khartoum
Electrical and Electronics Engineering

Arabic Language Modeling

Khalid N. Elmadani[1], Mukhtar Elgezouli[1]

[1]University of Khartoum, Electrical and Electronics Engineering

Abstract

Arabic, one of the six official languages of the United Nations, is the mother tongue of 300 millions people and yet it does not have a properly documented language model and benchmark dataset. A language model is used to build up a sequence of words, classes or phrases which are linguistically valid without any use of external knowledge. A list of probabilities is estimated from a large corpus to indicate the likelihood of linguistic events. This kind of models is useful in a large variety of research areas: speech recognition, optical character recognition, machine translation, spelling correction and much more. We here present a language model with two different approaches (LSTM and GRU) trained with text taken from wikipedia, perplexity and Bleu score are used to evaluate the Language model for each approach.

Data Prepossessing

- We used articles from latest Wikipedia arabic dataset (Arabic wiki dumps).
- We used wikiextractor to extract articles from the compressed file to xml files.
- Process-wikipedia Script with nltk tokenizer are used to convert xml files to a single csv file containing all articles and their lengths.
- Before splitting the data into training, validation and test sets we took only the articles which between 30 and 70 in length. So, we used only 212,146 Articles out of 907,705 (148,502 For training, 31,823 For validation, 31,821 For testing).

Models

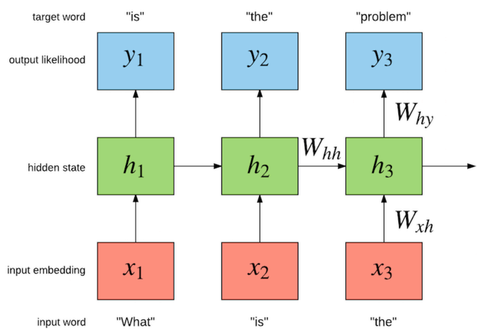

Figura 1: General Architecture of Language Models

- The Vectorizer goes through all the data making a vocabulary of words and indexes duals, a cutoff of (2) prohibits the words that occur less than two times from entering the vocabulary ,so the vocabulary starts at index 0 up to the number of words that occur more than two, the words are then sorted with the most occuring word coming first.
- The dataloader vectorize the batched input sequences then The embedding layer (first layer in all the models) vectorize each word in the sequence into an embedding vector (holding meaning of each word) using AraVec (an Arabic word embedding model).
- Two approaches are then available LSTM and GRU.
- For Both language models we used Adaptive Log Soft-max to reduce the training time.

Evaluation metrics

The goal is to assign a higher probability to grammatically correct and frequent sentences than those sentences which are rarely encountered or have some grammatical error. This can not be done by just comparing test set accuracy and loss for different models. On this we are using two evaluation metrics:

Perplexity

Perplexity is the multiplicative inverse of the probability assigned to the test set by the language model, normalized by the number of words in the test set.

$$PP(W) = P(w_1 w_2 ... w_N)^{-\frac{1}{N}}$$

$$= \sqrt[N]{\frac{1}{P(w_1 w_2 ... w_N)}}$$

Figura 2: Preplexity equation

Better language models will have lower perplexity values.

Blue score

The Bilingual Evaluation Understudy Score, or BLEU for short, is a metric for evaluating a generated sentence to a reference sentence. The approach works by counting matching n-grams in the candidate translation to n-grams in the reference text, where 1-gram or unigram would be each token and a bigram comparison would be each word pair. The comparison is made regardless of word order, A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0 for each word, normalized by the number of words in the test set.

Results

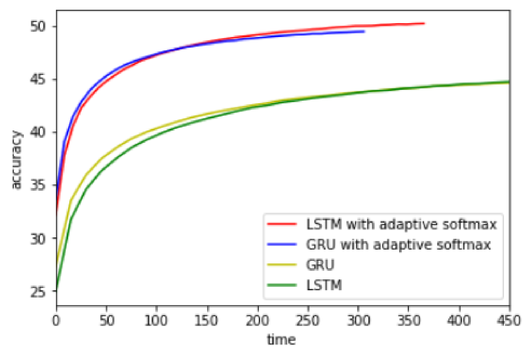

Figura 3: using softmax vs adaptive softmax

Model	perplexity	Blue Score
GRU	666	0.508
LSTM	639	0.519

Tabela 1: Comparing Results

Farther Work

We are aiming to produce a benchmark data-set for Arabic language by evaluating each sentence using the Perplexity and Bleu score (and other measures) of this language model, and try to improve or delete the sentences with bad measures manually.