Optimising word embeddings for recognised multilingual speech

SACAIR2020

Introduction

- South Africa is a multilingual society → require custom embedding training techniques.
- Smaller datasets of low-resource languages
- Code-switched text corpora are hard to find
- Most research use monolingual datasets.
- Monolingual datasets is not sufficient in a multilingual environment (Pratapa et al, 2018)
    - Syntactic structures
    - Cross-lingual semantic associations

- Different architectures perform similar when hyper-parameter settings are standardised. (Li et al, 2014)
- Each NLP task has peculiar characteristics and challenges
- How to optimise hyper-parameters to apply the model in the context of recognised multilingual speech?

Introduction

10. Two approaches to evaluate word embeddings:
    a. Intrinsic (e.g. similarity and analogy tasks)
        i. Widely used
        ii. Does not correlate to real-world applications
    b. Extrinsic
        i. Test performance on real-world applications.
11. There is no direct correlation between the performance of an embedding on intrinsic tasks and their performance on a real-world problem. (Chiu et al, 2016), (Schnabel et al, 2015), (Linzen, 2016), (Gladkova and Drozd, 2016)
12. Television and radio broadcasts can naturally be classified into categories:
    a. news, advertisements, sport, traffic and weather.

Hyper-parameters

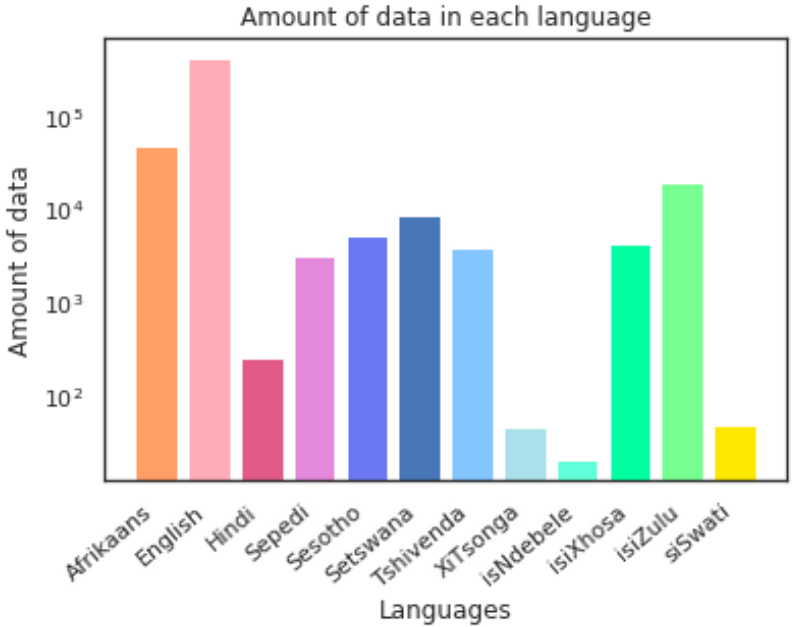
1. **Embedding size**
   a. Number of dimensions in the embedding layer.
   b. Related to, but usually greater than, vocabulary size.
2. **Window size**
   a. Number of context words around the target word.
   b. Default = 5.
   c. Can be computationally expensive.
3. **Training algorithm**
   a. Number of times the algorithm iterates over the training data.
   b. Default = 5.
   c. Can be computationally expensive.

Hyper-parameters

1. **Minimum count**
   a. Model ignores words that occur less than the minimum count.
2. **Learning rate**
   a. Between 0.05 and 0.5
   b. A higher learning rate over a longer time is beneficial for rare words but a learning rate decay creates better word
3. **Negative sampling**
   a. Cost function only calculated on a subset of input ->speeds up the process.
4. **Negative sampling distribution**
   a. Samples words in relation to their frequencies.
   b. ... low-frequency words more than high-frequency words.

Data

1. Speech-recognition outputs produced by Saigen on South African radio and television broadcasts obtained from Novus.
2. Recordings of 103 different South African radio and television stations.
3. English radio station data was used
    a. Includes code switching to other languages
    b. Useful representation of spoken South African English.
4. After pre-processing: 100 K words
    a. Training - 70 K
    b. Testing - 30 K



Data

1. As the level of code switching increases, the performance is expected to decrease (Gambäck and Das, 2016).
2. Multilingual Index (M-Index) quantifies the ratio of languages in the corpus based on the Gini-coefficient.
    a. Measures the degree in which a language is distributed in the dataset
    b. Does not indicate if the languages are integrated with each other

$$M - Index = \frac{1 - \sum_j p_j^2}{(k-1)\sum_j p_j^2} \quad (1)$$

where $k > 1$ is the number of languages, p_j is the number of words in a language over the total number of words in the dataset. j range over all the languages present in the dataset. The M-index is bounded between 0 (a monolingual dataset) and 1 (where each language in the dataset is represented equally)

Data

1. Integration index (I-index) complements the M-index
    a. Sum up the probability that there has been a switch.
    b. Does not require dividing the dataset into utterances or for it to contain computing weights.

the dataset into utterances or for it to contain computing weights. Given a dataset where each token has a language tag $\{l_i\}$ where i ranges from 1 to n, the size of the dataset:

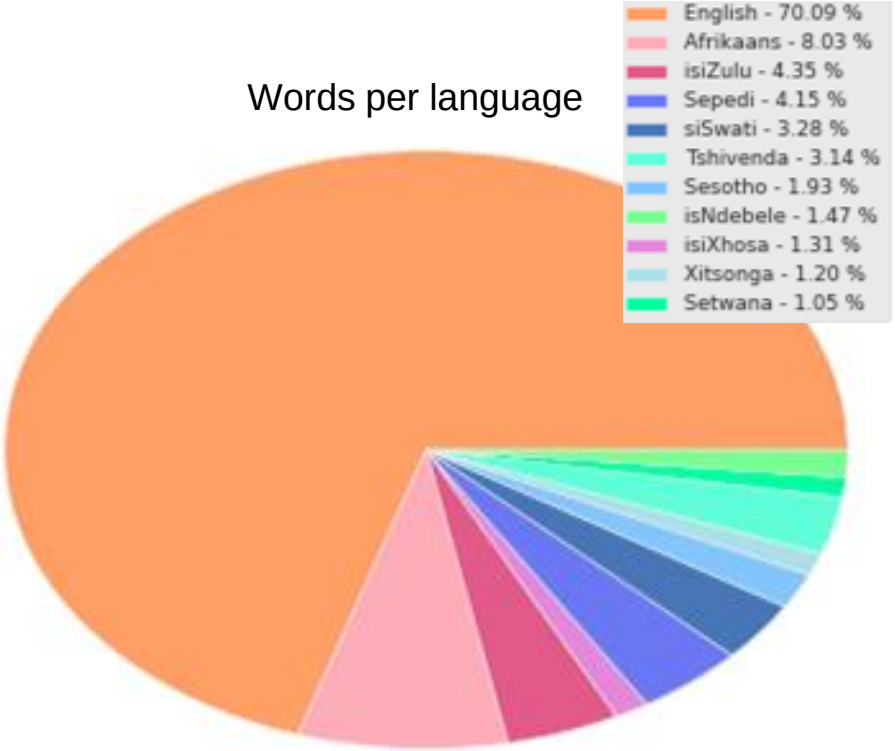
$$I - index = \frac{1}{n-1} \sum_{1 \leq i < i+1 \leq n} S(l_i, l_{i+1}) \qquad (2)$$

where $S(l_i, l_{i+1}) = 1$ if $l_i \neq l_{i+1}$ and 0 otherwise. The index is bounded between 0 (where no switching occurs) and 1 (if code switching occurs between each word pair.)

Data

1. To measure the M-index and the I-index:
    a. Tag each word in the corpus with their language.
    b. Whatlang: Word-level language identification
    c. Custom models for the South African languages
2. M-index: Distribution of the languages in the dataset
    a. M-Index = 0.32
    b. Indicate uneven representation of different languages
3. I-index: Number of times a language switch occurred in the data
    a. I-index = 0.299
    b. Suggesting a large amount of code switching.

Words per language



Models

1. CBOW -> predict a word, given the context
2. CBOW model performs better with news data, it is more stable and less data is needed (Mikolov et al, 2013 & Kim et al, 2019)

Model	**Dataset**
Various code-switched CBOW Word2Vec models	Code-switch dataset
BoW	Code-switch dataset
GloVe pre-trained	Wikipedia 2014 + Gigaword 5

Hyper-parameter selection

Table 1: The variable testing range for each hyper-parameter. Highlighted in red are the default values for each hyper-parameter.

Hyper-parameter	testing range
Embedding size	50, 100, 150, 200, 250, 300
Window size	2, 5, 10, 15, 30, 40, 50
Training time	2, 5, 10, 20, 40, 80, 160
Minimum count	0, 2, 4, 6
Learning rate	0,0025, 0.025, 0.25
Negative sampling	0, 5, 10, 15, 20
Negative sampling distribution	-0.5, -0.25, 0, 0.25, 0.5, 0.75, 1

Evaluation: Text classification task

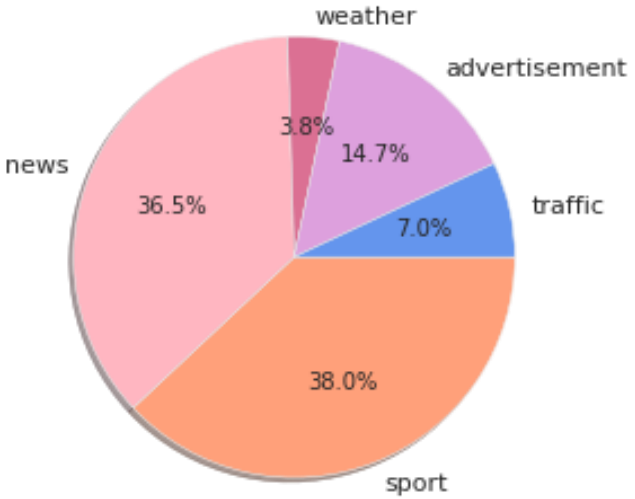
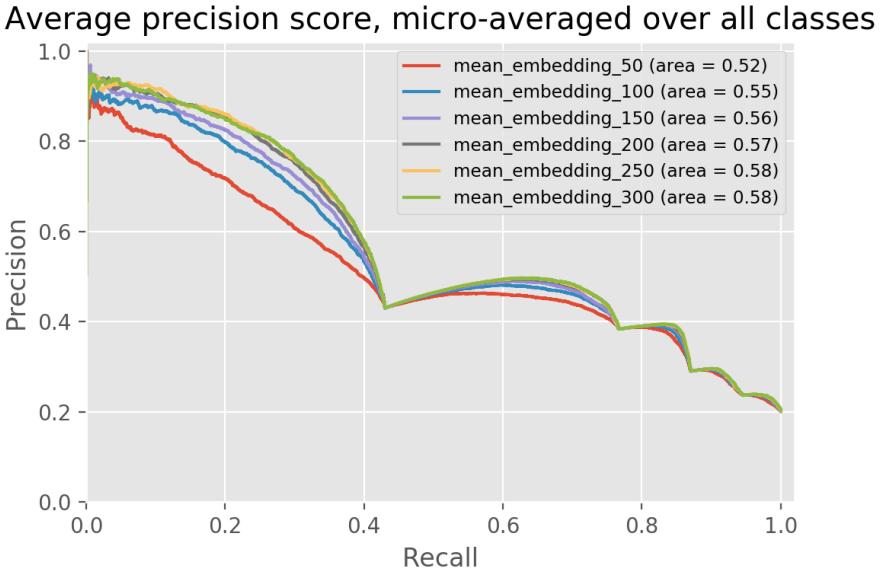
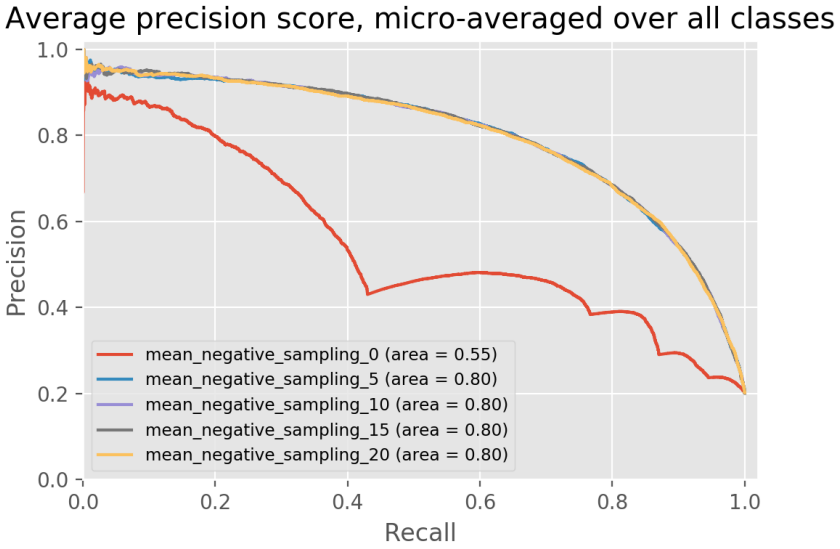


Table 2: Examples of sentences in the dataset tagged with their appropriate catagories.

Tag	Sentence:
News	in limpopo the body of the deceased was found next to the road in the bush buck ridge area last week police spokesperson leonard tladi says the mother appeared in court yesterday also on a charge of murder
Sport	the southern kings go in search of their second win of the season against conduct while the cheetahs take aim at...
Traffic	inbound slows down between sable road and the elevated freeway traffic lights are faltering retreat at prince george drive and military road
Advertisement	with our easy to use guide dstv keeps you up to date with latest episodes and action from us connect to your explorer just schedule recordings so you never miss download the app and keep the world at your fingertips
Weather	it's around twenty-one degrees in joburg good morning

Comparing the feature selection models



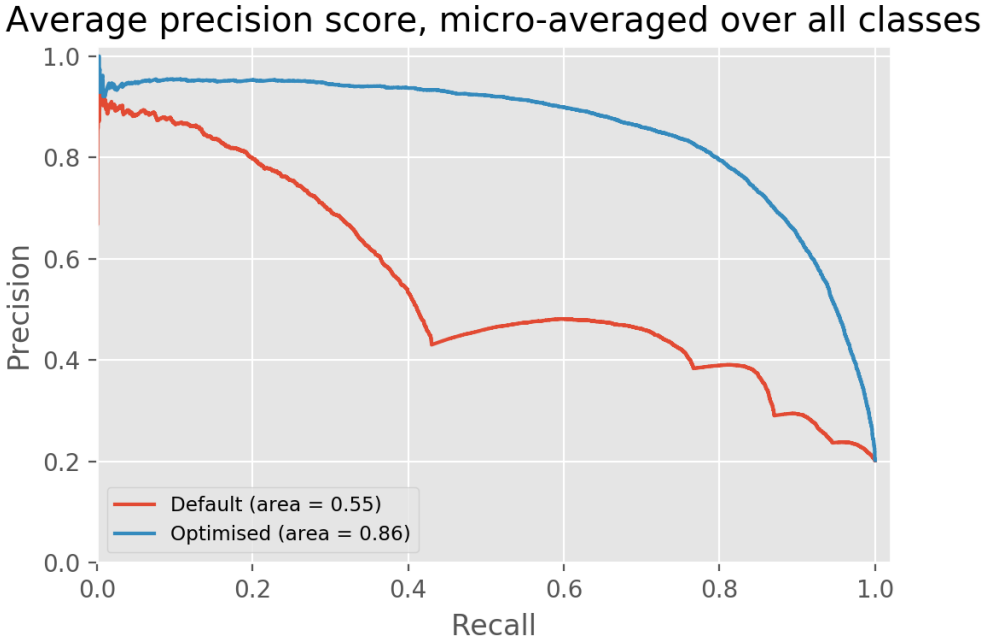


Comparing the interactions of multiple parameters

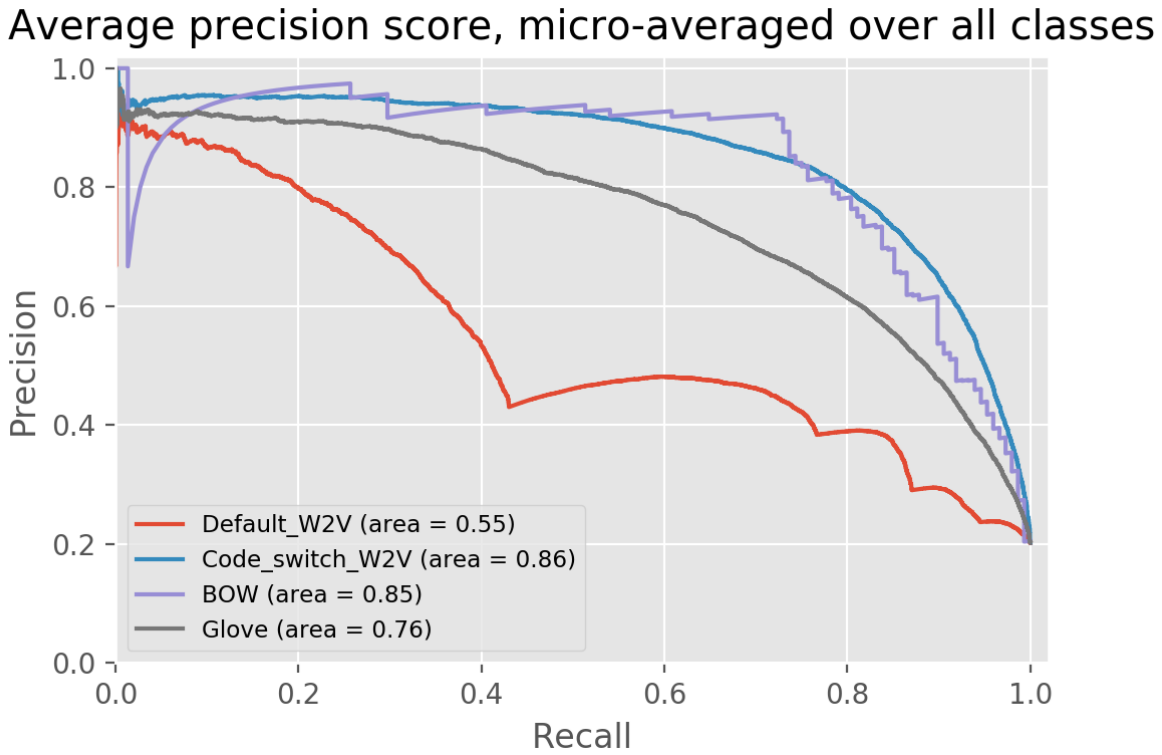
Negative sampling	Embedding size/training time/learning rate/ns distribution	Positive improvement
Negative sampling	Window size/minimum count	No improvement
Negative sampling = 5	Negative sampling distribution =0	2%
Negative sampling = 5	Training time =10 epochs	Optimal
Negative sampling	Learning rate	10%
Learning rate	Negative sampling distribution =0.75	0.0025
Learning rate	Negative sampling distribution =0	0.025
Window size => 30	Minimum count	0%
Window size	Minimum count	1%

Evaluating the code-switched model

Hyper-parameter	Default value	Optimised value
Embedding size	100	250
Window size	5	15
Training time	5	10
Minimum count	2	2
Learning rate	0.025	0.025
Negative sampling	0	5
Negative sampling distribution	0.75	0



Comparison between baseline and feature selection models



Conclusion

1. Word2Vec is standardised with default hyper-parameter values optimised in the original research where the embeddings were used to derive analogies for word pairs.
2. Each NLP task has its own characteristics and challenges.
3. Important parameters:
    a. Embedding size
    b. Negative sampling + appropriate distribution
4. Hyper-parameters combinations:
    a. Window size
    b. Training time
    c. Minimum count
5. These findings show:
    a. The importance of optimising certain hyper-parameters to fit the task
    b. The potential of embeddings to process recognised multilingual speech.