USING DEEP LEARNING TO IDENTIFY LANGUAGES IN SHORT TEXT

Jeanne Elizabeth Daniel

October 5, 2018

University of Stellenbosch

INTRODUCTION

INHERENT BIASES IN NATURAL LANGUAGE PROCESSING TOOLS

Majority of NLP research is limited to languages that are spoken by a majority population (e.g. English, Spanish, Mandarin)

- More incentives (recognition, leader boards) to research NLP tools for major languages
- More resources get developed and verified for major languages
- Even more resources get developed on top of these major language resources (more breakthroughs)
- Less access (e.g. World Wide Web, Research Centres) is given to native speakers of minor languages
- Speakers of low resource languages can create content in their native language, but is not always supported or language-tagged, resulting in fewer digital resources available for these languages

WHY DO WE NEED TO IDENTIFY LANGUAGES?

- Generate more annotated text through language tagging in metadata on websites and social media,
- Enable automatic machine translation,
- Helps chatbots know which language to respond in,
- Building web crawlers that can collect language-specific metadata.

Of the official languages, Facebook supports Afrikaans and English, Twitter supports only English. Google Translate supports English, Afrikaans, Zulu, Sesotho, and Xhosa.

NEURAL NETS FOR LANGUAGE IDENTIFICATION

Neural networks have shown promising results in identifying language, and even code-switching in short text. Using neural network architectures, we aim to develop a language identification tool for short text ($<$ 300 characters).

We perform our case study on the 11 official languages of South Africa: English, Afrikaans, Zulu, Xhosa, Southern Sotho, Tswana, Northern Sotho/Sesotho Sa Leboa, SiSwati, Ndebele, Venda, and Tsonga.

IMPLEMENTATION

MAIN OBJECTIVES

1. Acquire a reliable dataset to develop and test our model on.
2. Develop cleaning rules and vectorizing our text.
3. Experiment with different Neural Network Architectures.
4. Evaluate performance using confusion matrix and accuracy on different length bins.

DATA ACQUISITION

We make use of the NCHLT cleaned text corpora containing metadata of all 11 languages, obtained from North West University Resource Catalogue. (`https://rma.nwu.ac.za/index.php/resource-index.html`).

This contains a mix of different domains, including law documents, website information, news and radio transcipts, etc.

We constructed our dataset as follows:

- Split all documents into sentences
- Keep all sentences longer than 49 characters
- Drop duplicates (16.6% of original dataset)
- Initial class imbalance where English was almost 25% - halve the number of English texts
- Once-off randomly split into Train, Validate, and Test set (60:20:20)

This leaves us with 372,241 training samples, 123,700 validation samples, and 123,956 samples to test on.

Table: Breakdown of Class Distribution within Dataset

Language	**% of dataset** (as is)	**% of dataset** (post normalizing)
English	24.18	16.03
Zulu	11.32	14.58
Northern Sotho	9.38	11.38
Afrikaans	11.78	10.76
Sesotho	7.99	9.34
Xhosa	8.10	8.58
Setswana	5.33	6.73
Tsonga	5.74	6.40
SiSwati	6.26	6.26
Ndebele	5.77	5.38
Venda	4.14	4.56

DATA CLEANING

This includes:

- removing unwanted (non-ascii) characters
- replacing characters that are unique identifiers of a language with their closest cousin (alphabet) – they might not occur in the wild this way
- standardizing our length to 300 characters, by truncating or padding using white space

FEATURE REPRESENTATION

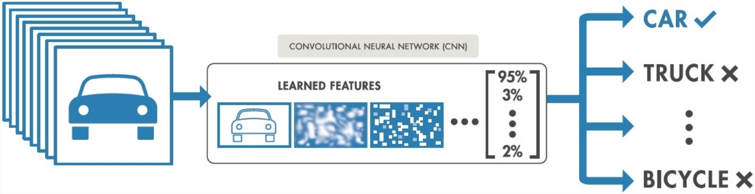
Feature Representation is done by mapping alphabet letters to their rank, we map each letter of the alphabet to its rank, starting with a $\rightarrow$ 1 … z $\rightarrow$ 26, and whitespace to 0:

For example, “bula faele” (‘open file’) will be transformed to

[2, 21, 12, 1, 0, 6, 1, 5, 12, 5]

CONVOLUTIONAL LAYERS

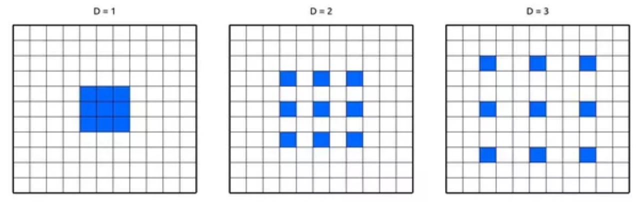
A biology-inspired variant of neural networks. The filters emulate the response of an individual neuron to visual stimuli. Convolutional filters are passed over the input, and pick up unique features, before feeding it forward.



DILATED CONVOLUTIONAL FILTERS

Dilated convolutional filters:

- “dilates” the size of the filter,
- can learn and retain more information for the same number of parameters
- increased validation accuracy by 2%



CREATING OUR WORD EMBEDDINGS

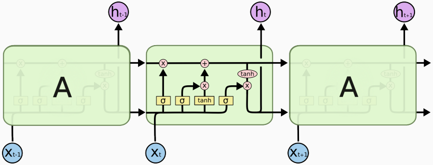
Word embedding spaces are:

- continuous vector spaces,
- capture relational information between topics and words,
- semantically similar words found very close to one another.

We make use of dilated convolutional filters to create our word embedding space. We don't use pooling layers, as we want to preserve as much information as possible.

LONG-SHORT TERM MEMORY.

LSTM stands for Long-Short Term Memory and extends the capacity of Recurrent Neural Networks by adding a “forget gate layer”. They are ideal for modeling temporal or sequential data, and help address the Vanishing/Exploding Gradient problem found with Recurrent Neural Networks.



We feed the word-embeddings created by the Convolutional layers into the LSTM layer for differentiation between languages.

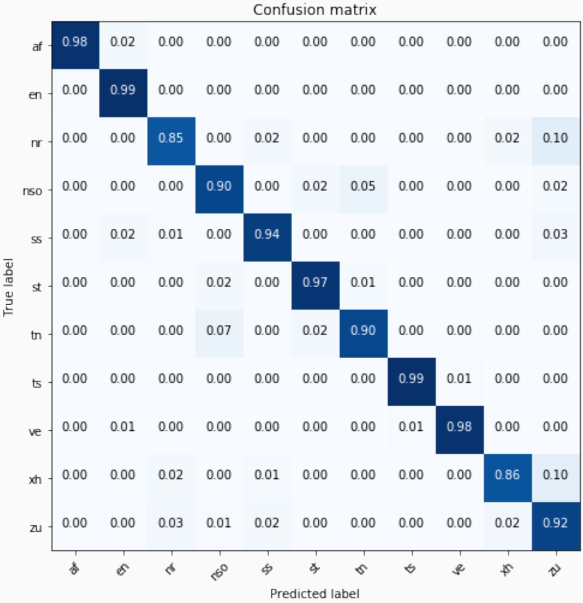
RESULTS

Table: Comparing for Train and Validation Accuracy within Bins of Lengths

Bins of Length	Train Acc %	Validation Acc %
$N < 100$	97.77	92.62
$100 \leq N < 200$	98.06	93.99
$200 \leq N < 300$	98.40	94.64
$N \geq 300$	98.52	95.46

Here we can see some over-fitting occurring, as the accuracy for the Training set is substantially higher than for the Validation set.

CONFUSION MATRIX



Confusion matrix

	af	en	nr	nso	ss	st	tn	ts	ve	xh	zu
af	0.98	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
en	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
nr	0.00	0.00	0.85	0.00	0.02	0.00	0.00	0.00	0.00	0.02	0.10
nso	0.00	0.00	0.00	0.90	0.00	0.02	0.05	0.00	0.00	0.00	0.02
ss	0.00	0.02	0.01	0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.03
st	0.00	0.00	0.00	0.02	0.00	0.97	0.01	0.00	0.00	0.00	0.00
tn	0.00	0.00	0.00	0.07	0.00	0.02	0.90	0.00	0.00	0.00	0.00
ts	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.01	0.00	0.00
ve	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.98	0.00	0.00
xh	0.00	0.00	0.02	0.00	0.01	0.00	0.00	0.00	0.00	0.86	0.10
zu	0.00	0.00	0.03	0.01	0.02	0.00	0.00	0.00	0.00	0.02	0.92

DISCUSSION

- There were some mislabeled items found in the NCHLT corpus itself, but the extent of this mislabelling was not investigated.
- Slight overfitting, when comparing the training vs validation accuracy
- Accuracy increases as length of text increases,
- Confusion between languages that are closely related, e.g. Zulu and Xhosa,
- Results are satisfactory, but training is costly.
- Could justify sacrificing some accuracy for speed, and opting for something like Multinomial Naive Bayes with n-grams.

THE END