Sentence Boundary Disambiguation using CBHG Model

Instructor: Prof. Lu Xiao

Kshitij Sankesara | Nishant Priyam | Rishabh Agrawal | Rucha Kadam

THE iSCHOOL Syracuse University

SYRACUSE UNIVERSITY · SUOS CULTORES SCIENTIA CORONAT · FOUNDED 1870

Introduction:

Sentence boundary disambiguation (SBD), also known as sentence tokenization, is a problem in Natural Language Processing (NLP) of deciding the begin and end points of a sentence. However, sentence boundary identification is challenging because punctuation marks can often become ambiguous while identifying. For example, a period may denote an abbreviation, decimal point, an ellipsis, or an email address – not the end of a sentence.

Dataset: We have used **Brown Corpus** for our data which we have obtained using nltk. The corpus contains 500 samples of English-language text, totaling around one million words. It is compiled from works published in the United States in 1961.

We have made use of Modified CBHG model, which was introduced first in the research paper called Tocatron: Towards End-to-End Speech Synthesis. It consists of reasonable number of hyperparameters and is a very powerful architecture.

Important Definitions:

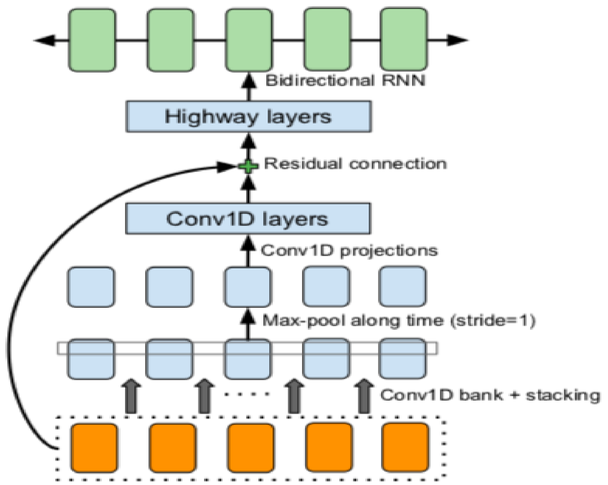
CNN: A convolutional neural network is a class of deep neural networks, most commonly applied to analyzing visual imagery. CNNs are regularized versions of multilayer perceptrons

Highway Network: A highway network is an approach to optimizing networks and increasing their depth. Highway networks use learned gating mechanisms to regulate information flow, inspired by LSTM recurrent neural networks.

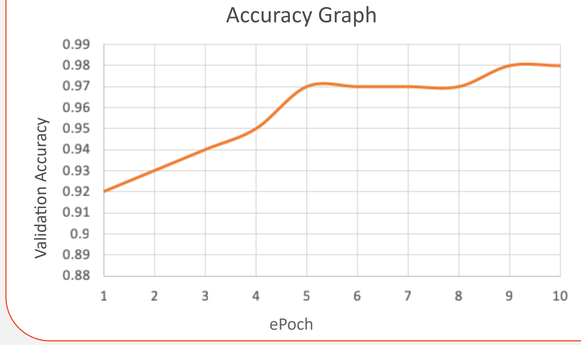
Bidirectional GRU: A gated recurrent unit (GRU) is part of a specific model of recurrent neural network that intends to use connections through a sequence of nodes to perform machine learning tasks associated with memory and clustering, for instance, in speech recognition.

LSTM: It is an artificial RNN architecture used in the field of deep learning. LSTM has feedback connections that make it a "general purpose computer". It can not only process single data points (such as images), but also entire sequences of data (such as speech or video).

CBHG Model Architecture



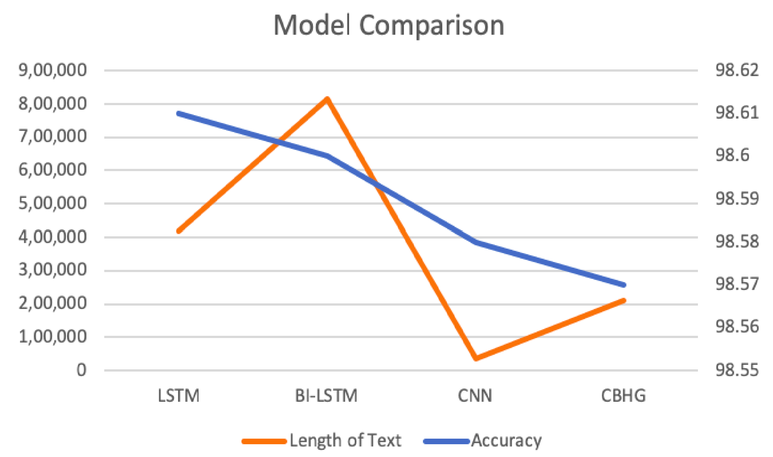
Model Graph:



Output:

```
Expected: At least he had the decency to blush she thought
Obtained: At least he had the decency to blush she thought

Expected: We caught the early train to New York
Obtained: We caught the early train to New York

Expected: Ran away on a black night with a lawful wedded man
Obtained: Ran a way on a black night with a lawful wedded man

Expected: He gave a short hard laugh and looked at her knowingly
Obtained: He gave a short hard laugh and looked at her knowingly

Final Accuracy = 208674/211699 = 0.9857
Baseline Accuracy = 166107/211699 = 0.7846
```

Results:



CBHG Model gives us approximately the same accuracy when compared to other models like LSTM, BI-LSTM and CNN.

Although, CBHG uses GRU with convolution bank which is expected to give better accuracy than LSTM when trained on larger dataset, LSTM gave marginally better accuracy.

References:

[1] Wang, Y., Skerry-Ryan, R., & Saurous, R. A. (2017, April 6). Tacotron: Towards End-To-End Speech Syn- Thesis. Retrieved from https://arxiv.org/pdf/1703.10135.pdf

[2] Schweter, S. (n.d.). General-Purpose Neural Networks for Sentence Boundary Detection. Retrieved from https://github.com/stefan-it/deep-eos

[3] Stanton, D., Wang, Y., & Skerry-Ryan, R. (2018, August 04). Predicting Expressive Speaking Style From Text In End-To-End Speech Synthesis. Retrieved from https://www.groundai.com/project/predicting-expressive-speaking-style-from-text-in-end-to-end-speech-synthesis/