# Quote Attribution Using Recurrent Neural Networks

Sayan Das
Computer Science and Engineering
SRM University, Kattankulathur
Chennai, India 603203
Email: sayandas@srmuniv.edu.in

***Abstract*—This paper proposes Defferent Neural Network Models for attributing dialogoues on Literary Texts. Quote Attribution, is the problem of tagging the dialogues with the speaker, is import for task like deanonymizing blogs, article or documents published illegally over world wide web, text mining models and as a media monitoring system. We used Recurrent Neural Models with defferent cells to establish the advantages and weak points among them.**

***Index Terms*—Natural Language Processing, Neural Networks, Machine Learning**

## I. Introduction

Conversation is the fundamental means for human interaction. Having back and forth interaction helps to understand the way one thinks, thus having a mutual understanding. Its the beauty of brain how after a prolonged observation of same person conversing it starts to infer patterns. These patterns can either be the repetition of words or unique grammatical structure. Here in this project I will try to replicate the same model using Natural Language Processing and Neural Networks. For this given a quote the model will try to interpret the speaker based purely on the language based approach, i.e. no visual and auditory data are provided. Model will try to interpret context by using clues like he said and she told to understand the gender of the speaker. This can be achieved by using Recurrent Neural Network (RNN) which stores the context upto a extent in the form of weights. Evaluation can be done by applying the predictive model to TV series episode scripts.

## II. Dataset

For the training and evaluation purpose I have used scripts of TV series particularly Seinfield. Script files have been collected by scraping *https://imsdb.com*. Several text preprocessing techniques are applied to clean the raw data.

## III. Text Preprocessing and Word Vector

Entire datasets have been cleaned by removing all the punctuations and bracket content, then converted to lower case and tokenised using NLTK tokeniser. We have used Stanford GloVe vector trained on *6 billion Gigaword5 + Wikipedia2014 Corpus* as the input layer for all the models. Given input word $w_i$ from the dialogue, $x_i$ be the one-hot row vector of the corresponding word from vocubulary(V) and $L \in R^{|V| \times d}$ be the embedding matrix, where $d$ is the dimension of the word vector. Then,

$$w_i = x_i \times L$$

## IV. Quote Vector

Given a quote $q$ consists of $\{w_1, w_2, w_3 ... w_n\}$ word vectors and the number of words $(n)$ in each quote varies. Each word vector $w_i$ can be represented by $\{f_1^i, f_2^i, f_3^i ... f_d^i\}$. here in this case the dimension of the quote is of $n \times d$, to reduce the dimension we have taken the summary of each quote and thus reducing the dimension to $d$.

$$\overrightarrow{q} = \frac{1}{n} \sum_{i=1}^{n} f_k^i, \forall k \in \{1, 2, 3...d\}$$

## V. Models

### A. Multi Layered Perceptron

## Appendix A
## Proof of the First Zonklar Equation

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

## Acknowledgment

The authors would like to thank...

## References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.